# Nonparametric inference for linear models via an analytic framework for the wild bootstrap

by

Katherine L. Burak

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

© Katherine L. Burak, 2023

# Abstract

Many standard approaches for conducting statistical inference on regression parameters rely heavily on parametric assumptions and asymptotic results. The wild bootstrap (Mammen, 1993) was developed as a nonparametric means to estimate a sampling distribution and is particularly useful when conducting statistical inference for linear models. With wide-reaching applications, the wild bootstrap can be used in a variety of settings where distributional assumptions are violated or are difficult to verify. While the wild bootstrap has attractive properties, as big-data becomes increasingly prevalent in society computationally intensive resampling schemes such as bootstrapping become less appealing and impractical.

In this work, an analytic framework for computing confidence regions and intervals in a variety of linear models is developed. The use of the concentration of measure phenomenon paired with the appealing properties of the wild bootstrap leads to a more computationally efficient, nonparametric way to perform statistical inference for regression parameters. The methodology is first introduced for the coefficients in least squares regression, and is then adapted to consider the more complex settings of ridge and LASSO regression. Lastly, this analytic approach is discussed in the context of generalized linear models, focusing on the case of overdispersion in Poisson regression.

# Preface

This thesis is an original work by Katherine Burak. It is composed of the research completed throughout my PhD at the University of Alberta in collaboration with my supervisor Dr. Adam Kashlak. The work discussed in Chapter 2 and a portion of Chapter 3 has been published as K. L. Burak and A. B. Kashlak, "Nonparametric confidence regions via the analytic wild bootstrap," *Canadian Journal of Statistics*. Articles based on Chapters 3, 4 and 5 are currently under preparation.

## Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Adam Kashlak. This endeavor would not have been possible without your guidance and feedback throughout the entire process. Your door was always open to talk about a problem or for some advice, and I am so grateful to have had the opportunity to work together.

To my parents, Lee and Jodi, words cannot express how thankful I am to have had your constant support throughout my academic journey. You have always believed in my dreams and pushed me to be the best version of myself. Thank you for being such amazing role models for me to look up to both as people and educators, and I am certain this achievement would not have been possible without you.

Lastly, I would like to extend my thanks to Dr. Brian Franczak and Dr. Linglong Kong for being a part of my advisory committee. I would also like to thank Dr. Brendan Pass and Dr. Liqun Diao for serving as members of my examination committee.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction and Overview

## 1.1 Motivation

When working with linear models, at times estimation and prediction may be the primary interest of the researcher. However, conducting statistical inference for regression coefficients such as hypothesis testing and confidence intervals is often a priority. While many standard, parametric approaches for inference exist, there are many cases when regression assumptions may be violated rendering these methods unsuitable. The wild bootstrap (Wu, 1986; Beran, 1986; Liu, 1988) was proposed as a means to estimate a sampling distribution under non-i.i.d. models. Mammen (1993) illustrates the asymptotic properties of the wild bootstrap estimator and shows that it correctly approximates the large sample distribution of the least squares estimator. Recently, the wild bootstrap has been applied to many regression problems such as quantile regression (Wang et al., 2018; Hagemann, 2017). Although the wild bootstrap may yield promising results, it is computationally inefficient. The wild bootstrap procedure involves perturbing the residuals by multiplying by a mean zero random variable a large number of times. This process is computationally costly and, when the sample size is large, can become quite inefficient.

This thesis explores nonparametric, analytic approaches to conducting statistical inference in various regression settings based on the wild bootstrap.

This involves analytically bounding the moments of the wild bootstrap estimator by applying a decoupling inequality (Kwapien, 1987; De la Pena and Giné, 2012). Sub-Gaussian bounds on the tail probability of the estimated regression parameters can then be established using concentration of measure (Ledoux, 2001; Boucheron et al., 2013). This approach only requires a small number of perturbations of the residuals, a fraction of what is demanded by the wild bootstrap, which saves a lot of computational time. Additionally, in response to the introduction of universal constants that arise from the decoupling inequality, an empirical beta transform Kashlak et al. (2022) is implemented in order to retain statistical power. This approach is primarily applied to compute confidence regions for regression parameters and linear contrasts, the latter allowing for the consideration of confidence intervals for individual regression coefficients. We explore these ideas in a variety of regression settings, starting with least squares regression, then transitioning into penalized regression and concluding with the consideration of generalized linear models.

## 1.2 Outline

The remainder of this thesis is organized as follows.

Chapter 2 discusses a variety of bootstrapping techniques including Mammen's wild bootstrap (Mammen, 1993) and reviews some relevant inequalities that are a basis for much of this work. As well, this chapter introduces the analytic wild bootstrap as a nonparametric way to build confidence regions and intervals for regression parameters. This chapter focuses on least squares regression and particularly emphasizes the importance of this approach in the case of heteroscedastic data.

In Chapter 3, we adapt these ideas to the penalized setting of ridge regression. The double bootstrap is implemented as a response to the problems that may arise when trying to bootstrap a biased estimator.

Chapter 4 extends our analytic approach to accommodate the coefficients in LASSO regression. We use a ridge-approximation to overcome the lack of closed-form expression available for the coefficients in LASSO regression.

Chapter 5 investigates applications of the wild bootstrap and ANWB to the class of generalized linear models. The value of these bootstrap approaches is demonstrated when presented with overdispersed count data.

Finally, Chapter 6 provides a conclusion and a discussion of potential future work in this area.

# Chapter 2

# Bootstrapping in Least Squares Regression

## 2.1 Introduction

Heteroscedastic errors are a major challenge that may arise when performing least squares regression. In particular, non-i.i.d. data poses a problem in the construction of confidence regions for regression parameters, where the classical parametric confidence regions derived using the $F$-distribution may not be appropriate. While nonparametric alternatives such as the wild bootstrap exist, they are not computationally efficient. In this chapter, an alternative, computationally efficient approach to constructing confidence regions for the regression parameters in least squares regression is proposed called the Analytic Wild Bootstrap (ANWB).

The ANWB is advantageous as it does not require either the assumption of normality or homogenous variance for the residuals and, in comparison to the wild bootstrap, is much more computationally efficient. For example, if we consider $B = 1000$ bootstrap replications for the wild bootstrap and $B = 10$ for the ANWB, using a quad-core processor a data set with dimensions $n = 100$ and $p = 5$ takes approximately 0.4 seconds CPU time for the ANWB and 24 seconds CPU time for the wild bootstrap. Moreover, a data set with larger dimensions like $n = 10000$ and $p = 100$ takes roughly 5 minutes CPU time for the ANWB and over 1.5 hours CPU time for the wild bootstrap.

The rest of this chapter is organized as follows. Section 2.2 describes how parametric confidence regions can be computed using the F-distribution. In Section 2.3, we review a variety of bootstrapping techniques in the context of least squares regression including Mammen's wild bootstrap (Mammen, 1993). Section 2.4 reviews the concentration of measure phenomenon as well as several relevant inequalities. Section 2.5 introduces the analytic wild bootstrap as a computationally efficient, nonparametric method for obtaining confidence regions for regression parameters. Simulation results are presented in Section 2.6, demonstrating the performance of the ANWB along with other methods for both homoscedastic and heteroscedastic data. Section 2.7 applies the methodology on a data set concerning the salaries and other aspects of professors from the Houston College of Medicine (Huang, 2017). A discussion of the results from the chapter is provided in Section 2.8. The work in this chapter has been published as Burak and Kashlak (2022).

## 2.2 Parametric Confidence Regions

Consider the linear model

$$Y = X\beta + \varepsilon,$$

where $Y$ is an $n$-dimensional vector of responses, $X$ is the $n \times p$ design matrix, $\beta$ is the $p$-dimensional parameter vector and $\varepsilon$ is the $n$-dimensional error term assumed to have mean zero and finite second moment. The least squares estimator $\hat{\beta}$ of $\beta$ is given by

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

In least squares regression where the residuals are i.i.d. from a normal distribution, it is convenient to consider the classical parametric confidence regions constructed using the $F$-distribution. Specifically, as outlined in Nickerson (1994), a $(1-\alpha)100\%$ confidence region for $\beta$ is defined by

$$(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)/p\text{MSE} \leq F_\alpha,$$

where MSE $= (Y'Y - \hat{\beta}'X'Y)/(n-p)$ and $F_\alpha = F(1-\alpha; p, n-p)$ is the $(1-\alpha)100$th percentile of the $F$-distribution. In appropriate settings, this approach is straightforward and sufficient for constructing confidence regions for $\beta$. However, when presented with heteroscedastic data or non-normal errors, it may not be suitable to consider this parametric construction that relies on the fact that $\varepsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2 I_n)$.

## 2.3 Bootstrapping

### 2.3.1 Residual resampling

In regression, a popular approach to bootstrapping is to use residual resampling introduced by Efron (1979). When residual resampling, the bootstrap model is

$$Y^* = X\hat{\beta} + \varepsilon^*,$$

where $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$ and $\varepsilon_i^*$ is obtained from sampling with replacement from the centered residuals. It is conventional to sample from the centered residuals as otherwise the bootstrap may fail (Freedman, 1981). The bootstrap estimate $\hat{\beta}^*$ of $\hat{\beta}$ is then

$$\hat{\beta}^* = (X'X)^{-1}X'Y^*.$$

In the presence of heteroscedastic data, however, residual resampling is not recommended as it eliminates the dependence between the error term and the explanatory variables (Freedman, 1981).

### 2.3.2 Pairs bootstrap

As an alternative to residual resampling, we can consider the pairs bootstrap (Freedman, 1981). The pairs bootstrap involves resampling vectors from the data as follows (Sartori, 2011),

1. From the observed data $(X, Y)$, sample with replacement from the pairs $(x_1, y_1), \ldots, (x_n, y_n)$ to get a bootstrap data set $(X^*, Y^*)$ consisting of

rows $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$.

2. Estimate bootstrapped regression coefficients as

$$\hat{\beta}^* = (X^{*'}X^*)^{-1}X^{*'}Y^*.$$

3. Repeat steps 1 and 2 $B$ times.

While convenient to use, the pairs bootstrap has some major disadvantages. Conventionally, the regression data $(x_i, y_i)$ is unbalanced and the pairs bootstrap does not take this into account (Wu, 1986). Since each bootstrap sample has a different design matrix $X^*$, the pairs bootstrap does not typically produce very accurate results (MacKinnon, 2006).

### 2.3.3 Wild bootstrap

Flachaire (1999) advises that the wild bootstrap should be used to overcome the drawbacks of the pairs bootstrap. The wild bootstrap Wu (1986); Beran (1986); Liu (1988) was introduced in order to accommodate non-i.i.d. models. The wild bootstrap model is

$$Y^* = X\hat{\beta} + \varepsilon^*,$$

where $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$ with $\varepsilon_i^* = \delta_i\hat{\varepsilon}_i$, where the $\delta_i$s are i.i.d. random variables with $\mathbb{E}[\delta_i] = 0$ and $Var[\delta_i] = 1$ (Mammen, 1993). The wild bootstrap estimator $\hat{\beta}^*$ of $\hat{\beta}$ is then

$$\begin{aligned}\hat{\beta}^* &= (X'X)^{-1}X'Y^* \\ &= (X'X)^{-1}X'X\hat{\beta} + (X'X)^{-1}X'\varepsilon^* \\ &= \hat{\beta} + (X'X)^{-1}X'\varepsilon^*.\end{aligned}$$

To gain more insight into how the wild bootstrap works, the following visualization is provided. Suppose we are working in the simple linear regression setting with $n = 5$ and a model is fit to the data as can be seen Figure 2.1.

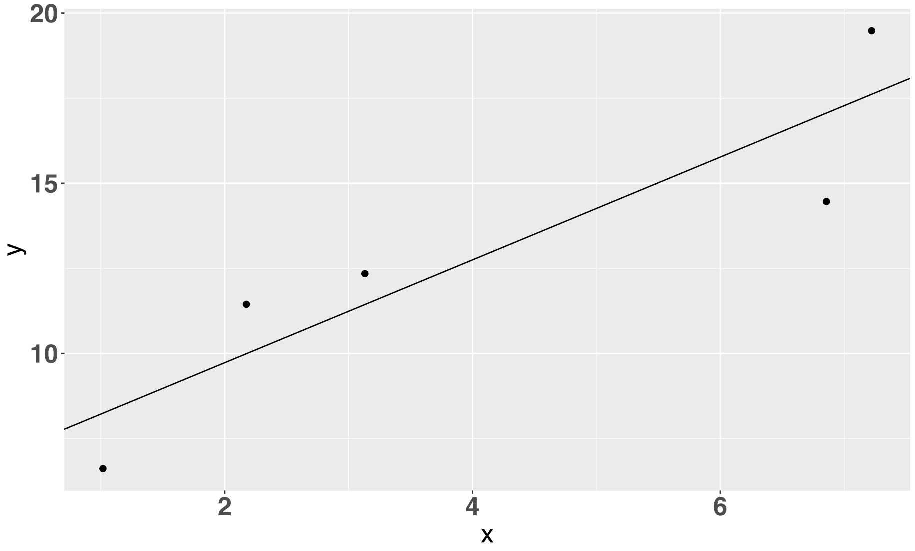

**Figure 2.1:** Simple linear regression with $n = 5$

We can perturb the response variable, $Y$, by multiplying the residuals by a Rademacher random variable, for example. These new points are denoted as $Y^*$ and are highlighted in blue in Figure 2.2. When $\delta_i = 1$, the original data point remains unchanged, whereas when $\delta_i = -1$, the observation is reflected across the fitted regression line.

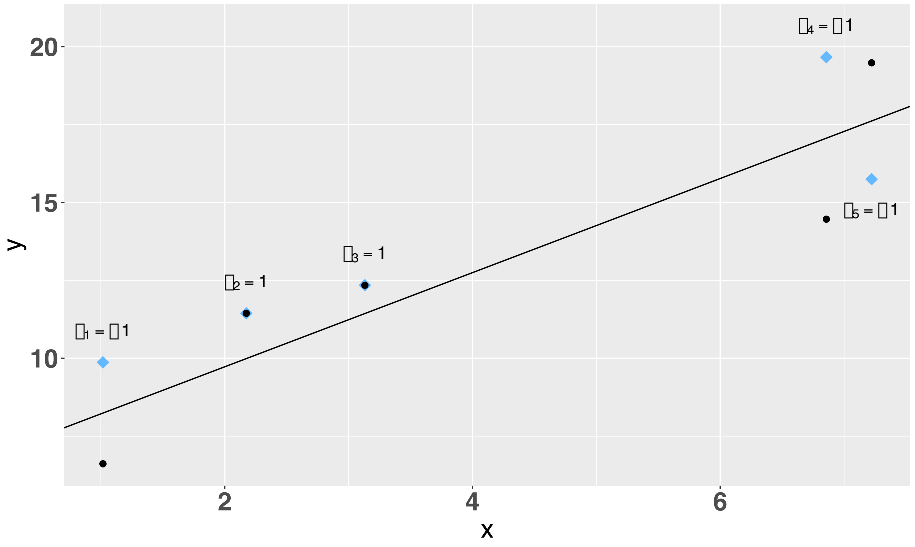

**Figure 2.2:** Simple linear regression with perturbed response variable

Finally, we can estimate the slope of the perturbed data points, $\hat{\beta}^*$, and are able to fit a new regression model as can be seen by the blue line in Figure 2.3.

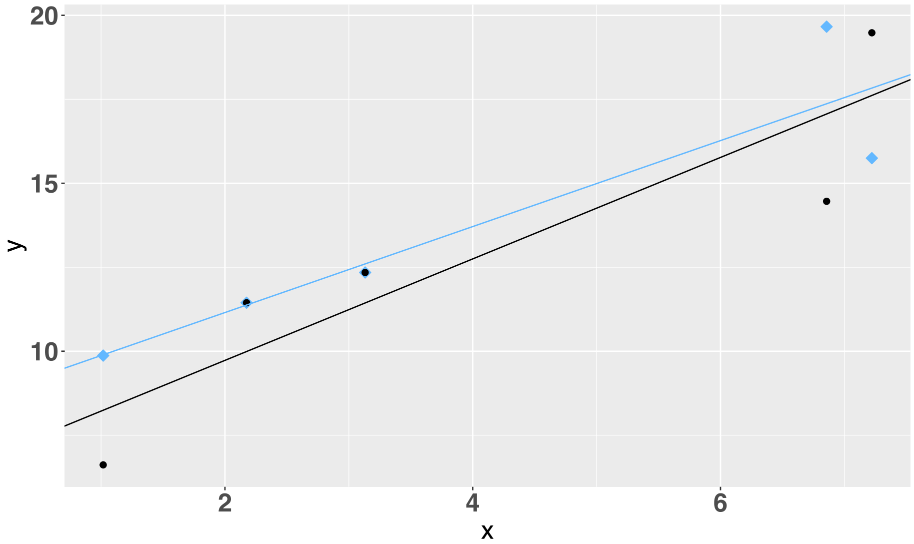

**Figure 2.3:** Wild bootstrap regression model

This process can be repeated a large number of times to estimate the sampling distribution of $\hat{\beta}$. As highlighted in Mammen (1993), the wild bootstrap estimator $\hat{\beta}^*$ exhibits asymptotic properties that agree with those of the least squares estimator $\hat{\beta}$. In particular, Mammen (1993) proved that, under certain conditions,

$$d_\infty\left(\mathcal{L}^*\left(\sqrt{n}c'\left(\hat{\beta}^* - \hat{\beta}\right)\right), \mathcal{L}\left(\sqrt{n}c'\left(\hat{\beta} - \beta\right)\right)\right) \xrightarrow{p} 0$$

for any $c \in \mathbb{R}^p$ such that $||c|| = 1$, where $d_\infty$ is the Kolmogorov distance and $\mathcal{L}^*(\dots)$ denotes the conditional distribution $\mathcal{L}(\dots|x_1, \dots, x_n, y_1, \dots, y_n)$. That is, the wild bootstrap consistently approximates the behaviour of least squares estimates. Thus, we are able to utilize the wild bootstrap estimator to construct a $(1-\alpha)100\%$ confidence region for $\beta$ by taking the $(1-\alpha)$ quantile of $(\hat{\beta}^* - \hat{\beta})'X'X(\hat{\beta}^* - \hat{\beta})$.

There are several choices for the distribution of the $\delta_i$s, the only restriction being that they have mean 0 and variance 1. For instance, Mammen (1993) proposed a distribution based on the golden ratio,

$$\delta_i = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } 1-(\sqrt{5}+1)/(2\sqrt{5}). \end{cases}$$

Another natural choice would be the standard normal distribution. As it has been shown to give better results (Flachaire, 2005), in this thesis we will use the Rademacher distribution, such that

$$\delta_i = \begin{cases} -1 & \text{with probability } 1/2, \\ 1 & \text{with probability } 1/2. \end{cases}$$

Although the wild bootstrap is appealing from a theoretical standpoint, in practice, as the dimensions of the data increase, the wild bootstrap becomes computationally inefficient as it requires a large number of replications. This makes it impractical when dealing with big data.

## 2.4 Overview of Relevant Inequalities

### 2.4.1 Concentration inequalities

Concentration of measure is a phenomenon that is widely applicable to many theoretical and applied areas of mathematics. Talagrand (1996) defines concentration of measure as the idea that “a random variable that depends (in a “smooth” way) on the influence of many independent variables (but not too much on any of them) is essentially constant.” In other words, this means that although a random variable may take on a large number of possible values, those likely to be observed are concentrated in a very small range (Dubhashi and Panconesi, 2009).

Concentration inequalities typically bound the difference between the fluctuations of functions of independent random variables around their expected values (Boucheron et al., 2013). The law of large numbers is one simple example of a concentration inequality. In what follows, some famous results and instances of concentration inequalities are provided as discussed in Boucheron et al. (2013).

Perhaps one of the most basic concentration inequalities is Markov's inequality.

**Theorem 2.4.1** (Markov's inequality)**.** *Let $X$ be a nonnegative random variable. Then for all $t > 0$,*

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Chebyshev's inequality can then be derived applying Markov's inequality.

**Theorem 2.4.2** (Chebyshev's inequality)**.** *Let $X$ be a real-valued random variable. Then,*

$$P(|X - \mathbb{E}[X]| \geq t) \leq \frac{Var[X]}{t^2}.$$

To get a sharper bound, Markov's inequality can be applied to derive Chernoff bounds. From Markov's inequality, for every $\lambda \geq 0$,

$$P(X \geq t) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}].$$

Selecting $\lambda$ to minimize the upper bound, we have that

$$P(X \geq t) \leq \inf_{\lambda \geq 0} e^{-\lambda t}\mathbb{E}[e^{\lambda X}].$$

For example, for $X \sim \mathcal{N}(0, \sigma^2)$, this approach can be used to show that $P(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$.

This notion of exponentially decaying bounds leads us to the idea of sub-Gaussian random variables. Define $\psi_X(\lambda) = \log \mathrm{E}[e^{\lambda X}]$ as the logarithmic moment generating function of $X$. A centered random variable $X$ is sub-Gaussian with variance factor $v$ if

$$\psi_X(\lambda) \leq \frac{\lambda^2 v}{2} \text{ for every } \lambda \in \mathbb{R}.$$

We can characterize sub-Gaussian random variables in terms of their tail probabilities. Hence, if $X$ is a centered sub-Gaussian random variable with variance factor $v$, then for all $t > 0$,

$$P(X \geq t) \leq e^{-\frac{t^2}{2v}}.$$

Using Chernoff's inequality, the case of sums of bounded independent real-valued random variables can be considered via Hoeffding's inequality.

**Theorem 2.4.3** (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i$ takes values in $[a_i, b_i]$ almost surely for all $i \leq n$. Let*

$$S = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]).$$

*Then for every $t \geq 0$,*

$$P(S \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Hoeffding's lemma establishes a sub-Gaussian bound on the tail probability of $S$, but we may require inequalities with sharper bounds such as Bennett's and Bernstein's inequalities.

**Theorem 2.4.4** (Bennett's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with finite variance such that $X_i \leq b$ for some $b \geq 0$ almost surely for all $i \leq n$. Let*

$$S = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])$$

*and $v = \sum_{i=1}^{n} \mathbb{E}[X_i^2]$. If we write $\phi(u) = e^u - u - 1$ for $u \in \mathbb{R}$, then, for all $\lambda > 0$,*

$$\psi_S(\lambda) \leq nlog\left(1 + \frac{v}{nb^2}\phi(b\lambda)\right) \leq \frac{v}{b^2}\phi(b\lambda),$$

*and for any $t > 0$,*

$$P(S \geq t) \leq \exp\left(-\frac{v}{b^2} h\left(\frac{bt}{v}\right)\right),$$

*where $h(u) = (1 + u)log(1 + u) - u$ for $u > 0$.*

**Theorem 2.4.5** (Bernstein's inequality)**.** *Under the same settings as Bennett's inequality,*

$$P(S \geq t) \leq \exp\left(-\frac{t^2}{2(v + bt/3)}\right).$$

The type of concentration depends on the relationship between the size of $t$ in relation to $v/b$. If $v$ is relatively large, Bennett's and Bernstein's inequality yield a sub-Gaussian concentration. However, if $t >> v/b$, the result is a sub-exponential type of inequality.

### 2.4.2 Decoupling inequalities

The notion of decoupling refers to the idea of comparing a sum of dependent random variables to one of independent random variables, thus "decoupling" them by removing the dependence. The study of decoupling inequalities started emerging in the 1980's (see Kwapien (1987); McConnell and Taqqu (1986)), but decoupling still is being applied in more modern works (see Alves and Sapozhnikov (2019); Makarychev and Sviridenko (2018)). A thorough review of decoupling inequalities can be found in De la Pena and Giné (2012). The decoupling result primarily applied in this thesis is Corollary 3 from Kwapien (1987). By introducing some independence to dependent random variables, decoupling is a technique that can greatly simplify the derivations of certain mathematical results (De la Pena and Giné, 2012).

### 2.4.3 Khintchine inequalities

Although the result has been around for a relatively long time, Khintchine's inequality (Khintchine, 1923) still has a wide variety of applications in areas such as mathematics, probability theory and even computer science (De et al., 2016; Floret and Matos, 1995; Buchholz, 2001; Nguyen et al., 2009). The Khintchine inequality is useful because it allows for the comparison of $L_p$ and $L_2$ norms of sums of weighted independent Rademacher random variables (Spektor, 2016). In particular, Khintchine's inequality is a pertinent tool when analyzing decoupling inequalities (De la Pena and Giné, 2012). Khintchine's inequality is provided below as described in Kashlak et al. (2022) and Garling (2007).

**Theorem 2.4.6** (Khintchine's inequality)**.** *For any $p \in (0, \infty)$, there exist positive finite constants $A_p$ and $B_p$ such that for any sequence $x_1, \ldots, x_n \in \mathbb{R}$,*

$$A_p^p ||x||_2^p \leq \mathbb{E} \left| \sum_{i=1}^{n} \delta_i x_i \right|^p \leq B_p^p ||x||_2^p,$$

*where $\delta_1, \ldots, \delta_n$ are iid Rademacher random variables, $A_{2p} = [(2p)!/2^p p!]^{-1/2p}$ and $B_{2p} = [(2p)!/2^p p!]^{1/2p}$.*

In this thesis, we are primarily interested in the upper bound of Khintchine's inequality. Other variations of Khintchine's inequality have been investigated such as the restricted Khintchine inequality which allows for the consideration of random vectors with dependent coordinates (Spektor, 2016).

## 2.5 Analytic Wild Bootstrap

### 2.5.1 Confidence regions

In this section, a nonparametric approach to constructing confidence regions for $\beta$ called the Analytic Wild Bootstrap (ANWB) is introduced for both least squares and ridge regression that is much less computationally expensive than the wild bootstrap and is still capable of handling heteroscedastic data. Note that we work under the setting where $n > p$ so that we can consider $(X'X)^{-1}$.

**Theorem 2.5.1.** *For a linear model $Y = X\beta + \varepsilon$ with independent, not necessarily identically distributed errors $\varepsilon$, denote $\hat{\beta}$ as the least squares estimator of $\beta$ and the residuals as $\hat{\varepsilon}_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\beta} x_i$. Consider the wild bootstrap model $Y^* = X\hat{\beta} + \varepsilon^*$, where $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$ with $\varepsilon_i^* = \delta_i \hat{\varepsilon}_i$ and the $\delta_i$s are i.i.d. from a symmetric distribution such that $\mathbb{E}[\delta_i] = 0$ and $Var[\delta_i] = 1$, we denote $\hat{\beta}^*$ as the wild bootstrap estimator of $\hat{\beta}$. Furthermore, assume that for*

*$p \geq 2$ and for a constant $B_p$,*

$$\left(\mathbb{E}_\delta \left|\sum_{i=1}^{n} \delta_i \hat{\varepsilon}_i\right|^p\right)^{1/p} \leq B_p \left(\mathbb{E}_\delta \left|\sum_{i=1}^{n} \delta_i \hat{\varepsilon}_i\right|^2\right)^{1/2}$$

*for each sequence $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$. Then, for $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,*

$$P\left((\hat{\beta}^* - \hat{\beta})' X' X (\hat{\beta}^* - \hat{\beta}) \leq -\log(\alpha^*)\, 4C \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}\right) \geq 1 - \alpha,$$

*where $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ is the regularized incomplete beta function, $\theta_1$ and $\theta_2$ are fixed unknown constants, $C$ is a universal constant and $h_{ij}$ is the $(i,j)^{th}$ element of $H_X = X(X'X)^{-1}X'$.*

*Proof.* In order to construct a confidence region for $\beta$, consider the following quadratic form.

$$\begin{aligned}
(\hat{\beta}^* - \hat{\beta})' X' X (\hat{\beta}^* - \hat{\beta}) &= \varepsilon^{*'} X (X'X)^{-1} X' X (X'X)^{-1} X' \varepsilon^* \\
&= \varepsilon^{*'} H_X \varepsilon^* \\
&= \sum_{i,j=1}^{n} h_{ij} \varepsilon_i^* \varepsilon_j^* \\
&= \sum_{i,j=1}^{n} h_{ij} \delta_i \delta_j \hat{\varepsilon}_i \hat{\varepsilon}_j \\
&= T^2.
\end{aligned}$$

This proof follows the work done in Kashlak et al. (2022). As $T^2$ is a degree 2 polynomial chaos, it follows from Kwapien (1987) that

$$\begin{aligned}
\left[\mathbb{E}_\delta \left|\sum_{i,j=1}^{n} h_{ij} \delta_i \delta_j \hat{\varepsilon}_i \hat{\varepsilon}_j\right|^p\right]^{1/p} &\leq B_p^2 C \left[\mathbb{E}_\delta \left|\sum_{i,j=1}^{n} h_{ij} \delta_i \delta_j \hat{\varepsilon}_i \hat{\varepsilon}_j\right|^2\right]^{1/2} \\
&= B_p^2 C \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2},
\end{aligned}$$

where $B_{2p}$ is determined by the distribution of the $\delta_i$s. In the case of Rademacher or Gaussian random variables we take $B_{2p} = [(2p)!/2^p p!]^{1/2p}$. Hence,

$$\mathbb{E}_\delta \left|T\right|^{2p} \leq B_p^{2p} C^p \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{p/2}.$$

Now, let $T'$ be an independent copy of $T$. We adapt the moment bounds into a tail bound as follows.

$$\begin{aligned}\mathbb{E}_\delta \left[e^{\lambda T}\right] &\leq \mathbb{E}_\delta \left[e^{\lambda(T-T')}\right] \\ &= \sum_{p=0}^{\infty} \frac{\lambda^p}{p!} \mathbb{E}_\delta \left[T - T'\right]^p \\ &\leq \sum_{p=0}^{\infty} \frac{\lambda^{2p}}{(2p)!} \mathbb{E}_\delta \left[T - T'\right]^{2p}. \qquad (2.1)\end{aligned}$$

Note that $\mathbb{E}_\delta \left[T - T'\right]^{2p} = \mathbb{E}_\delta \left[2T/2 - 2T'/2\right]^{2p} \leq 2^{2p-1}\left(\mathbb{E}_\delta \left[T\right]^{2p} + \mathbb{E}_\delta \left[T\right]^{2p}\right)$ $= 2^{2p}\mathbb{E}[T]^{2p}$. Updating $C$, Equation (2.1) becomes

$$\begin{aligned}\mathbb{E}_\delta \left[e^{\lambda T}\right] &\leq \sum_{p=0}^{\infty} \frac{\lambda^{2p} 2^{2p}}{(2p)!} \mathbb{E}_\delta \left[T\right]^{2p} \\ &\leq \sum_{p=0}^{\infty} \frac{\lambda^{2p} 2^{2p}}{(2p)!} B_p^{2p} C^p \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{p/2} \\ &= \sum_{p=0}^{\infty} \frac{\lambda^{2p} 2^{2p} C^p \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{p/2}}{(2p)!} \left(\left[\frac{p!}{2^{p/2}(p/2)!}\right]^{1/p}\right)^{2p} \\ &= \sum_{p=0}^{\infty} \frac{\lambda^{2p} 2^{2p} C^p \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{p/2}}{(2p)!} \frac{(p!)^2}{2^p (p/2)!^2} \\ &= \sum_{p=0}^{\infty} \lambda^{2p} 2^p C^p \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{p/2} \frac{(p!)^2}{(2p)!(p/2)!^2} \\ &\leq \left[e/\sqrt{\pi} + e/\pi\right] \sum_{p=0}^{\infty} \frac{\left(\lambda^2 C \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}\right)^p}{p!},\end{aligned}$$

where the last inequality uses the result that $\frac{(p!)^2}{(2p)!(p/2)!^2} \leq [e/\sqrt{\pi} + e/\pi]/p!$ (Watson, 1959). Thus,

$$P(T > t) \leq \left[e/\sqrt{\pi} + e/\pi\right] e^{-t^2/4C\left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}}. \tag{2.2}$$

From this bound, we can compute a $(1-\alpha)100\%$ confidence region for $\beta$ as

$$(\hat{\beta}^* - \hat{\beta})' X' X (\hat{\beta}^* - \hat{\beta}) \leq -\log\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}\right) 4C \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}.$$

Applying the empirical beta transformation (Kashlak et al., 2022) with simulated parameters $\hat{\theta}_1$ and $\hat{\theta}_2$, we construct the confidence region based on the adjusted level

$$\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \hat{\theta}_1, \hat{\theta}_2\right).$$

Therefore, a $(1-\alpha)100\%$ confidence region for $\beta$ can be computed as

$$(\hat{\beta}^* - \hat{\beta})' X' X (\hat{\beta}^* - \hat{\beta}) \leq -\log(\alpha^*) 4C \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2},$$

such that

$$P\left((\hat{\beta}^* - \hat{\beta})' X' X (\hat{\beta}^* - \hat{\beta}) \leq -\log(\alpha^*) 4C \left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}\right) \geq 1 - \alpha,$$

as desired.

□

The calculation of the so-called universal constant, $C$, introduced by the decoupling inequality depend on the distribution of the $\delta_i$s. Following Corollary 3 of Kwapien (1987), in the case of Rademacher random variables with $d = 2$, $C = d^{3d}/d! = 2^6/2! = 2^5$ and is updated to $C = 2^6$ as we obtain a sub-Gaussian bound. The term "updating" refers to the universal constant's absorption of other constant terms as we compute this bound. If we consider the $\delta_i$s to be standard normal random variables, the universal constant can be improved since $(\delta_1, \dots, \delta_n)$ is a 2-stable symmetric random vector when each $\delta_i \sim \mathcal{N}(0,1)$. Hence, with $d = p = 2$, $C$ can be computed as $C = (d^{d/p}/d!)/d^{d(1/p-1)} = (2^{2/2}/2!)/2^{2(1/2-1)} = 2$, similarly updated as $2^2 = 4$ when obtaining a sub-Gaussian bound.

As noted by Kashlak et al. (2022), the main drawback of the use of decoupling inequalities is a loss of statistical power. The universal constants inflate the bounds of the respective confidence regions, resulting in regions that are too conservative to be meaningfully interpreted. This is why we consider the use of the empirical beta transformation (Kashlak et al., 2022) in order to regain statistical power. As described in Theorem 2.4 of Kashlak et al. (2022), the bounds on the tail probabilities derived in the proof of Theorem 2.5.1,

$$\left[e/\sqrt{\pi} + e/\pi\right] e^{-t^2/4C\left[\sum_{i,j=1}^{n} h_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}},$$

where $t^2$ is a realization of $T^2 = (\hat{\beta}^* - \hat{\beta})' X' X (\hat{\beta}^* - \hat{\beta})$, follow a Beta$(\theta_1, \theta_2)$ distribution with unknown parameters $\theta_1$ and $\theta_2$ instead of a Uniform[0,1] distribution. Consequently, the empirical beta transform allows for the retention of statistical power by adjusting the bounds on the tail probabilities of $T$ via the incomplete beta function, where $\theta_1$ and $\theta_2$ are simulated using their respective method of moments estimators as is described in Algorithm 1.

Additionally, as mentioned in the proof of Theorem 2.5.1, $B_{2p}$ is determined by the choice of distribution for the $\delta_i$s. In the case of a Rademacher distribution, $B_{2p} = [(2p)!/2^p p!]^{1/2p}$ (Garling, 2007). This would also be the case for the Gaussian distribution as there is an equivalence between the two distributions (Ledoux and Talagrand, 2013). Other symmetric distributions satisfying a Khintchine-type inequality could be considered such as symmetric discrete uniform random variables (Havrilla and Tkocz, 2020), but $B_{2p}$ and the resulting confidence region would need to be adjusted accordingly.

**Algorithm 1** The Empirical Beta Transform for the ANWB

```
Choose $B > 1$, the number of bootstrap replicates to simulate—e.g. $B = 10$.

Draw $\delta_1, \dots, \delta_B$ uniformly at random.
Compute $B$ tail probabilities from Equation (2.2).

Find the method of moments estimators for $\theta_1$ and $\theta_2$ from the beta distribution.
    Estimate first and second central moments of the $B$ tail probabilities by $\bar{x}$ and $s^2$, the sample mean and variance.
    Estimate $\hat{\theta}_1 = \bar{x}^2(1-\bar{x})/s^2 - \bar{x}$.
    Estimate $\hat{\theta}_2 = [\bar{x}(1-\bar{x})/s^2 - 1][1-\bar{x}]$.

Construct confidence region based on the adjusted level
    $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \hat{\theta}_1, \hat{\theta}_2\right)$.
```

### 2.5.2 Linear contrasts

One may also be interested in deriving analytic confidence intervals for linear contrasts. Thus, in order to consider the linear contrast $c'(\hat{\beta}^* - \hat{\beta})$, we introduce the following theorem.

**Theorem 2.5.2.** *Under the same settings as Theorem 2.5.1, let $c \in \mathbb{R}^p$ be a fixed contrast vector and $v = (v_1, \dots, v_n)' = c'(X'X)^{-1}X'$. Furthermore, assume that for a constant $B_p$,*

$$\left[\mathbb{E}_\delta \left|\sum_{i=1}^n v_i \delta_i \hat{\varepsilon}_i\right|^p\right]^{1/p} \leq B_p^2 \left[\sum_{i=1}^n v_i^2 \hat{\varepsilon}_i^2\right]^{1/2}$$

*for each sequence $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n \in \mathbb{R}$. Then, for $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,*

$$P\left(\left|c'(\hat{\beta}^* - \hat{\beta})\right| \leq \left[-\log(\alpha^*)\, 4 \left[\sum_{i=1}^n v_i^2 \hat{\varepsilon}_i^2\right]^{1/2}\right]^{1/2}\right) \geq 1 - \alpha,$$

*where* $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$*,* $I(\alpha; \theta_1, \theta_2)$ *is the regularized incomplete beta function and* $\theta_1$*,* $\theta_2$ *are fixed unknown constants.*

*Proof.* Consider the linear contrast

$$\begin{aligned} c'(\hat{\beta}^* - \hat{\beta}) &= c'(X'X)^{-1}X'\varepsilon^* \\ &= v'\varepsilon^* \\ &= \sum_{i=1}^{n} v_i \varepsilon_i^* \\ &= \sum_{i=1}^{n} v_i \delta_i \hat{\varepsilon}_i \\ &= T, \end{aligned}$$

where $v = (v_1, \ldots, v_n)' = c'(X'X)^{-1}X'$. As $T$ is a function of the $\delta_i$s, we can apply the Khintchine inequality (Khintchine, 1923; Garling, 2007) to establish a bound on the moments of $T$, such that

$$\begin{aligned} \left[\mathbb{E}_\delta \left|\sum_{i=1}^{n} v_i \delta_i \hat{\varepsilon}_i\right|^p\right]^{1/p} &\leq B_p^2 \left[\mathbb{E}_\delta \left|\sum_{i=1}^{n} v_i \delta_i \hat{\varepsilon}_i\right|^2\right]^{1/2} \\ &= B_p^2 \left[\sum_{i=1}^{n} v_i^2 \hat{\varepsilon}_i^2\right]^{1/2}. \end{aligned}$$

Following the proof of Theorem 2.5.1, we adapt the moment bounds into the following tail bound,

$$P(|T| > t) \leq \left[e/\sqrt{\pi} + e/\pi\right] e^{-t^2/4\left[\sum_{i=1}^{n} v_i^2 \hat{\varepsilon}_i^2\right]^{1/2}}.$$

Similarly, we construct a confidence interval based on the adjusted level

$$\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \hat{\theta}_1, \hat{\theta}_2\right).$$

Therefore, a $(1-\alpha)100\%$ confidence interval for $T$ can be computed as

$$\left| c'(\hat{\beta}^* - \hat{\beta}) \right| \leq \left[ -\log(\alpha^*)\, 4 \left[ \sum_{i=1}^{n} v_i^2 \hat{\varepsilon}_i^2 \right]^{1/2} \right]^{1/2},$$

such that

$$P\left( \left| c'(\hat{\beta}^* - \hat{\beta}) \right| \leq \left[ -\log(\alpha^*)\, 4 \left[ \sum_{i=1}^{n} v_i^2 \hat{\varepsilon}_i^2 \right]^{1/2} \right]^{1/2} \right) \geq 1 - \alpha.$$

□

The consideration of linear contrasts enables the construction of individual confidence intervals for regression coefficients. For example, to construct a confidence interval for $\beta_j$, take $c$ as a vector of zeros with 1 in the $j^{th}$ entry. Consequently, we can conduct hypothesis tests for regression coefficients as follows. For $j = 1, \ldots, p$, we can test the hypothesis at the level $1 - \alpha$

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0,$$

by rejecting $H_0$ when 0 is within the corresponding two-sided $(1-\alpha)100\%$ confidence interval for $\beta_j$.

## 2.6 Simulations

In this section, we present simulation results comparing the coverage probabilities of confidence regions generated by the ANWB, the classical parametric approach based on the $F$-distribution (denoted as "parametric" in figures and tables) and the wild bootstrap. All simulations are replicated $r = 100$ times.

### 2.6.1 Least squares regression

We simulated our data from the following linear model

$$Y = X\beta + \varepsilon,$$

where $n \in \{100, 500, 1000\}$, $p = 5$, $\varepsilon \sim \mathcal{N}(0, 3)$, $X$ is generated from Uniform[0,5] and $\beta$ is generated from Uniform[0,3]. We take the $\delta_i$s to be i.i.d. Rademacher random variables with $P(\delta_i = 1) = P(\delta_i = -1) = 1/2$. For the wild bootstrap, we use $B = 1000$ bootstrap replications and construct a $(1 - \alpha)100\%$ confidence region for $\beta$ by taking the $(1 - \alpha)$ quantile of $(\hat{\beta}^* - \hat{\beta})' X' X (\hat{\beta}^* - \hat{\beta})$. For the ANWB, we use $B = 10$ replications as seen in Kashlak, Myroshnychenko and Spektor (2020). Additionally, we consider heteroscedastic data simulated with $\varepsilon \sim \mathcal{N}(0, \sigma_i^2)$, where $\sigma_i^2 = x_{i1}^2$ (Flachaire, 2005). Simulation results are presented for $\alpha = 0.05$.

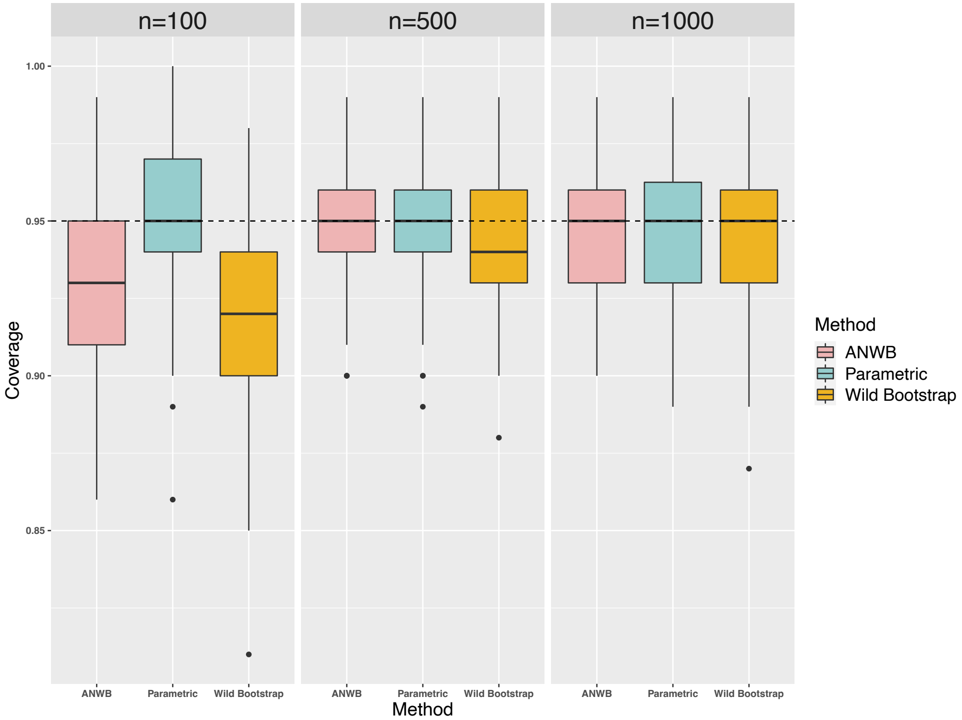

**Figure 2.4:** Coverage of 95% ANWB, parametric and wild bootstrap confidence regions for homoscedastic data

**Table 2.1:** Mean and standard deviation of the coverages of 95% ANWB, parametric and wild bootstrap confidence regions for homoscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| ANWB | 0.928 (0.028) | 0.95 (0.02) | 0.95 (0.02) |
| Parametric | 0.951 (0.022) | 0.95 (0.02) | 0.949 (0.022) |
| Wild Bootstrap | 0.917 (0.028) | 0.942 (0.021) | 0.945 (0.023) |

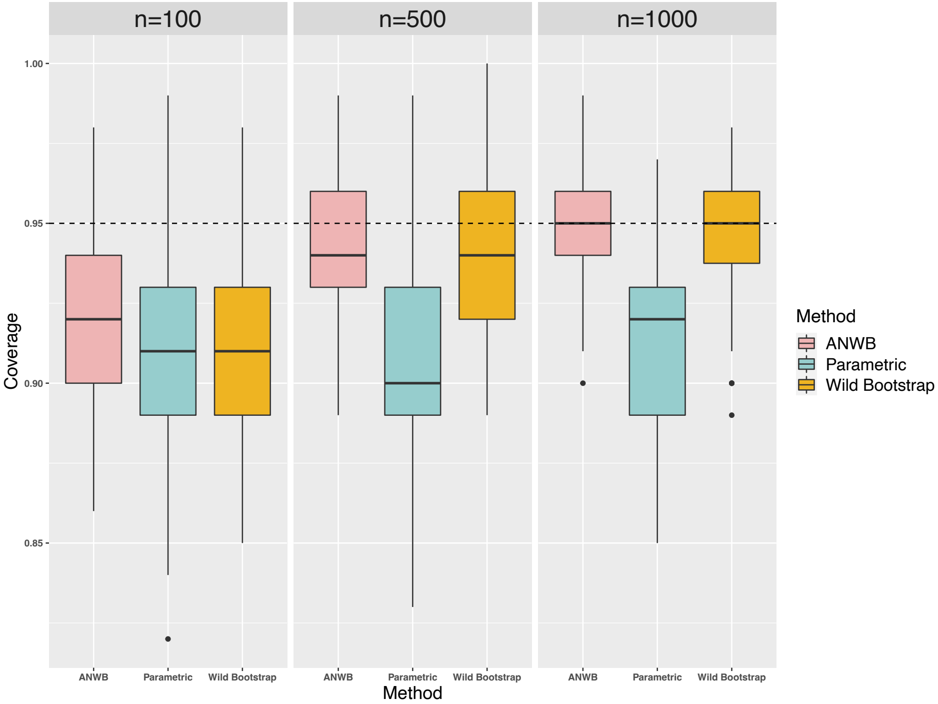

**Figure 2.5:** Coverage of 95% ANWB, parametric and wild bootstrap confidence regions for heteroscedastic data

**Table 2.2:** Mean and standard deviation of the coverages of 95% ANWB, parametric and wild bootstrap confidence regions for heteroscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| ANWB | 0.92 (0.026) | 0.945 (0.023) | 0.95 (0.02) |
| Parametric | 0.907 (0.03) | 0.906 (0.028) | 0.912 (0.027) |
| Wild Bootstrap | 0.907 (0.029) | 0.94 (0.023) | 0.948 (0.021) |

As can be seen in Figure 2.4 and Table 2.1, as $n$ increases the ANWB confidence regions attain the correct coverage. When $n$ is not sufficiently large, both the ANWB and wild bootstrap confidence regions are slightly too liberal. As is expected, in the case when the errors are i.i.d. from a normal distribution, the parametric confidence regions based on the $F$-distribution attain the desired coverage of 95%. In the case of heteroscedastic data presented in Figure 2.5 and Table 2.2, not surprisingly, the classical parametric confidence regions do not achieve the correct coverage. The ANWB and wild bootstrap perform well under heteroscedastic conditions as $n$ increases, with the analytic approach being much more efficient from a computational standpoint. Ultimately, the ANWB performs optimally in the case of heteroscedastic data, while taking a fraction of the computational time of the wild bootstrap.

### 2.6.2 Big data

Simulation results are presented for big data with $n = 10000$ and $p = 100$. The simulation settings are the same as in the previous section, with the exception of using $B = 100$ bootstrap replications for the wild bootstrap.

We can see in the case of homoscedastic errors as in Figure 2.6 and Table 2.3, as the dimensions of the data are increased the ANWB and parametric approaches exhibit quite similar behaviour. However, in both cases, the wild bootstrap confidence regions are too liberal. Additionally, in Figure 2.7 and Table 2.4 where the errors are heteroscedastic, the ANWB confidence regions outperform the other two methods. These results are fairly consistent to what we found in lower dimensions. As we look at higher-dimensional data, the computational advantages of the ANWB over the wild bootstrap become increasingly apparent. For larger data sets, the wild bootstrap becomes quite computationally inefficient which is why less bootstrap replications are considered for this simulation, consequently resulting in undercoverage by the wild bootstrap confidence regions. In order to utilize $B = 1000$ bootstrap replications over $r = 100$ total replications, the simulation for the wild bootstrap would require approximately 150 hours CPU time for a quad-core processor, which does not seem feasible. Thus, a researcher may be faced with the choice of considering the wild bootstrap with a low number of bootstrap replications or the ANWB. As using the wild bootstrap with a low number of bootstrap replications may result in undercoverage, the ANWB is preferable in higher-dimensional situations.

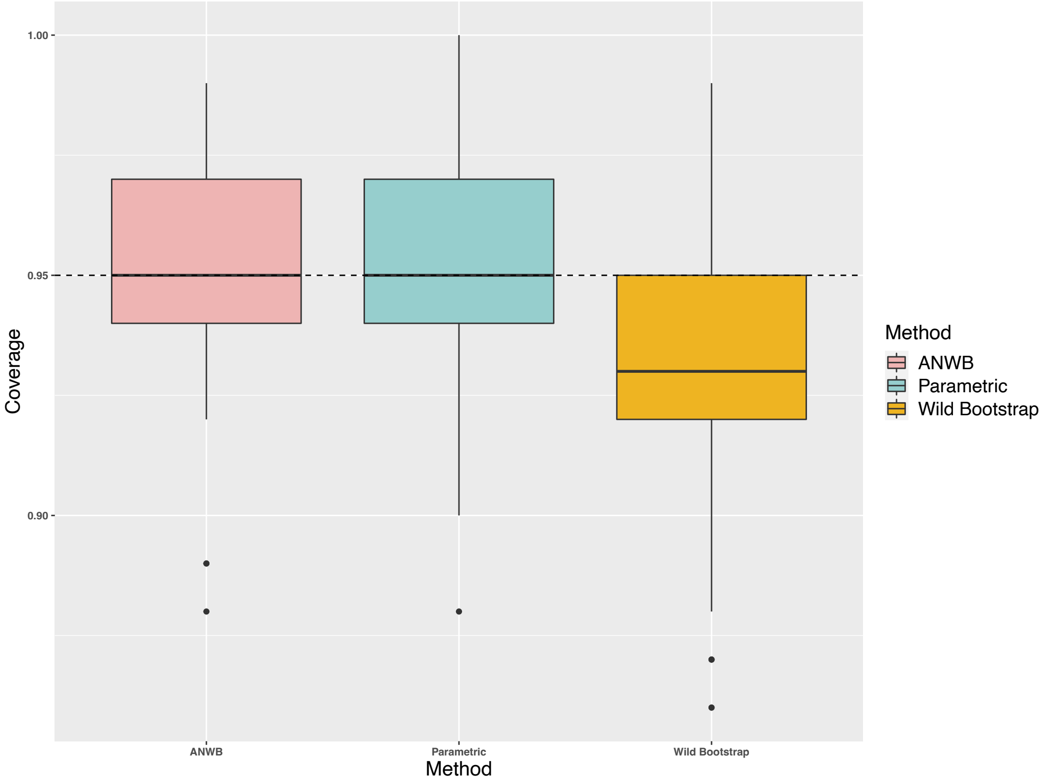

**Figure 2.6:** Coverage of 95% ANWB, parametric and wild bootstrap confidence regions for homoscedastic data

**Table 2.3:** Mean and standard deviation of the coverages of 95% ANWB, parametric and wild bootstrap confidence regions for homoscedastic data

| Method | Coverage |
|---|---|
| ANWB | 0.954 (0.023) |
| Parametric | 0.953 (0.022) |
| Wild Bootstrap | 0.931 (0.027) |

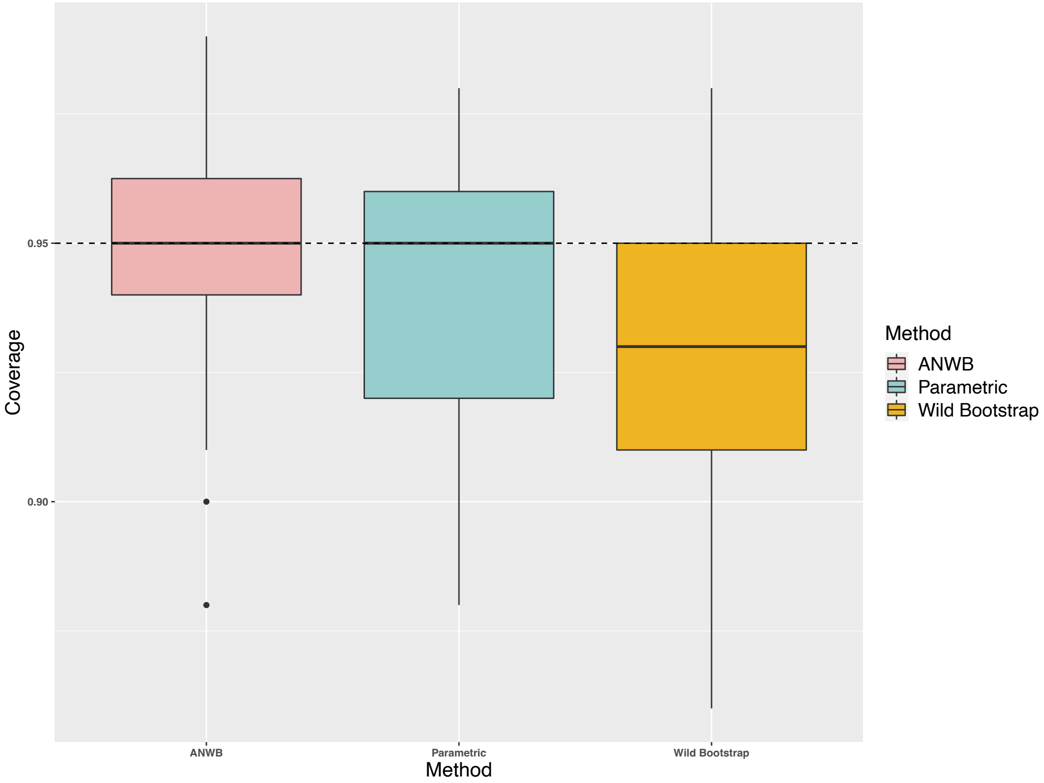

**Figure 2.7:** Coverage of 95% ANWB, parametric and wild bootstrap confidence regions for heteroscedastic data

**Table 2.4:** Mean and standard deviation of the coverages of 95% ANWB, parametric and wild bootstrap confidence regions for heteroscedastic data

| Method | Coverage |
|---|---|
| ANWB | 0.95 (0.02) |
| Parametric | 0.942 (0.024) |
| Wild Bootstrap | 0.929 (0.025) |

### 2.6.3 Linear contrasts

We simulated our data from the following linear model

$$Y = X\beta + \varepsilon,$$

where $n \in \{100, 500, 1000\}$, $p = 10$, $\varepsilon \sim \mathcal{N}(0, 3)$, $X$ is generated from Uniform[0,5] and $\beta$ is generated from Uniform[0,3]. We take the $\delta_i$s to be i.i.d. Rademacher random variables with $P(\delta_i = 1) = P(\delta_i = -1) = 1/2$. Here, we illustrate the construction of confidence intervals for $\beta_1$ by using a contrast vector of $c = (1, 0, 0, 0, 0)$ as is outlined in Theorem 2.5.2. Simulation results are presented for $\alpha = 0.05$.

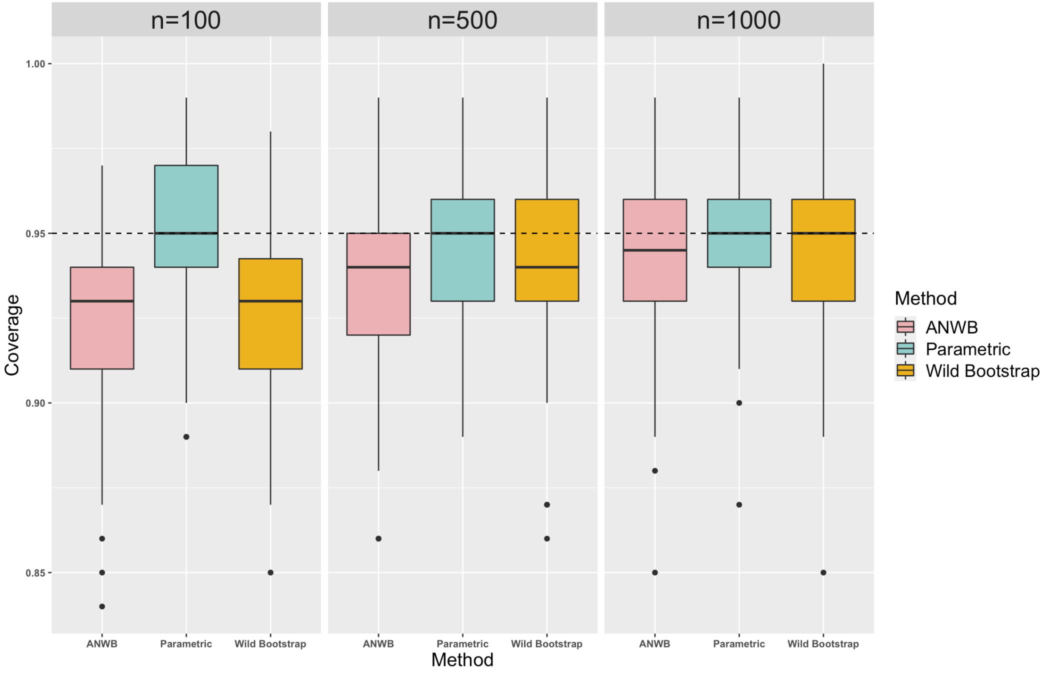

**Figure 2.8:** Coverage of 95% ANWB, parametric and wild bootstrap confidence intervals for homoscedastic data

**Table 2.5:** Mean and standard deviation of the coverages of 95% ANWB, parametric and wild bootstrap confidence intervals for homoscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| ANWB | 0.924 (0.028) | 0.938 (0.026) | 0.943 (0.0246) |
| Parametric | 0.950 (0.022) | 0.947 (0.022) | 0.950 (0.022) |
| Wild Bootstrap | 0.927 (0.029) | 0.942 (0.024) | 0.945 (0.024) |

## 2.7 Data Example

In this section, we consider the analysis of a data set concerned with attributes of professors at the Houston College of Medicine (Huang, 2017). This data set was collected in order to highlight gender discrimination of female faculty members regarding wage and faculty position. The data is comprised of $n = 261$ observations and $p = 10$ variables. Consider the following model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_2^2 + \beta_6 X_3^2 + \beta_7 X_2 \times X_3,$$

where $Y$ represents the scaled faculty member's salaries in 1995, $X_1$ represents sex (female or male), $X_2$ represents the publication rate (# publications on cv)/(# years between CV date and MD date), $X_3$ represents the number of years since obtaining an MD and $X_4$ represents the board certification status (not board certified or board certified). Looking at the residual plot in Figure 2.9, it appears the assumption of homoscedasticity is not satisfied. The estimated regression equation is

$$\hat{Y} = 0.07X_1 - 1.77X_2 + 0.96X_3 + 0.11X_4 + 1.35X_2^2 - 0.20X_3^2 - 0.54X_2 \times X_3,$$

with a test statistic of 113.7 on 7 and 254 degrees of freedom and corresponding p-value $< 2.2 \times 10^{-16}$. For ANWB confidence regions, we take the $\delta_i$s to be i.i.d. Rademacher random variables and $C = 2^6$. ANWB, parametric and wild bootstrap confidence regions are plotted for 6 regression coefficients below in Figure 2.10; all three methods produced similar confidence regions, although the analytic approach tends to yield more variable regions in practice.

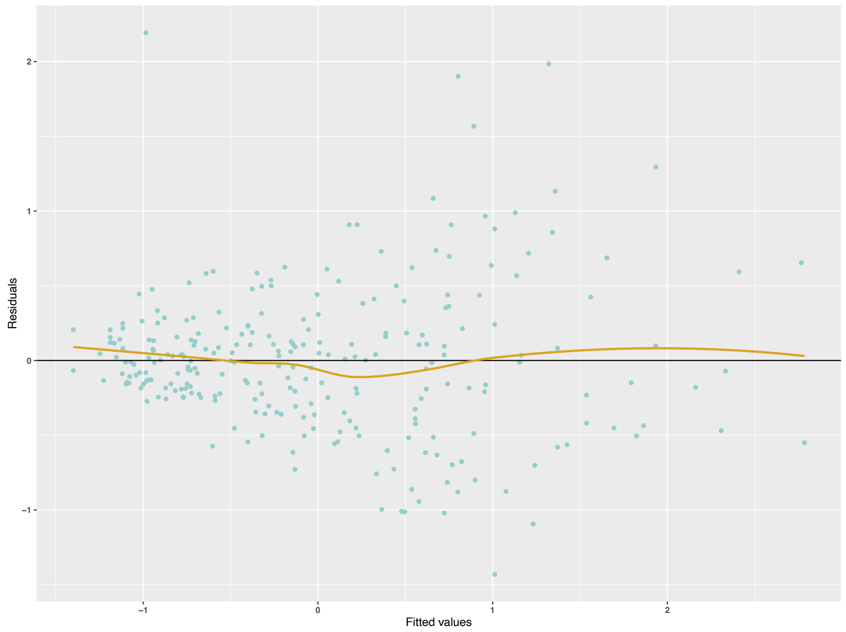

**Figure 2.9:** Residuals vs. fitted values for salary data set

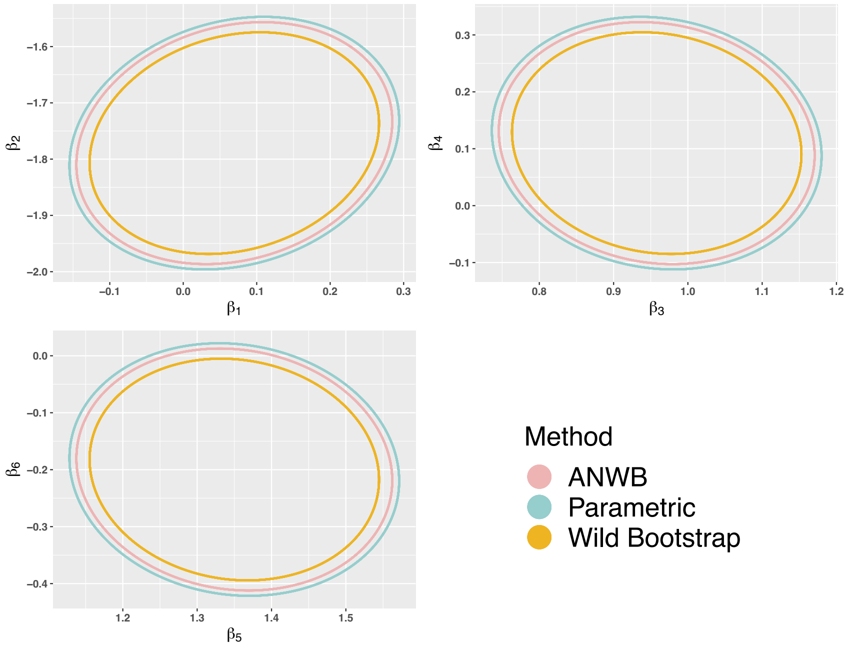

**Figure 2.10:** 95% ANWB, parametric and wild bootstrap confidence regions for salary data set

## 2.8 Discussion

It was demonstrated that the finite sample performance of the ANWB confidence regions approach the correct level as $n$ becomes sufficiently large. When the dimensions of the data are relatively high and the errors are i.i.d. from a normal distribution, the ANWB and parametric confidence regions yield similar results. In the case of heteroscedastic data where classical parametric confidence regions may not be appropriate, the ANWB maintains good results and is much more computationally efficient than the wild bootstrap, especially in higher-dimensional settings. The ANWB is a more efficient, nonparametric alternative to constructing confidence regions for big data. In Chapters 3 and 4, we investigate the extension of the analytic wild bootstrap to the penalized regression setting.

# Chapter 3

# Analytic Bootstrap for Ridge Regression

## 3.1 Introduction

Ridge regression was first introduced by Hoerl and Kennard (1970) and is a form of penalized regression that was proposed to accommodate data that exhibits multicollinearity. The ridge regression estimator $\hat{\beta}(\lambda)$ is a shrinkage estimator with an $L_2$ penalty such that, as defined in Efron and Hastie (2016),

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ ||Y - X\beta||^2 + \lambda \, ||\beta||^2 \right\}.$$

A researcher may be interested in obtaining confidence regions or intervals for the coefficients in ridge regression. However, $\hat{\beta}(\lambda)$ is biased so typical parametric approaches may not be suitable. Moreover, as is mentioned in Section 3.2, applying the bootstrap to ridge regression can be misleading and although it may appear to work in certain situations, it is not advisable.

In this chapter we discuss a viable approach for bootstrapping in ridge regression using the double bootstrap and apply the analytic wild bootstrap in this setting as well. The rest of this chapter is organized as follows. Section 3.2 addresses the single bootstrap applied to ridge regression and discusses why it is not necessarily appropriate. Section 3.3 introduces the double bootstrap and fast double bootstrap as feasible alternatives to bootstrapping in ridge regression. Furthermore, in Section 3.4 the AFDB is developed as an analytic solution to computing confidence regions, drastically reducing computational time. Simulations are presented in Section 3.5. Section 3.6 highlights the methodologies on a data set about the energy efficiency of buildings. In Section 3.7, the various approaches are discussed and summarized.

## 3.2 The Single Bootstrap

### 3.2.1 Wild bootstrap

In this section, we will outline the wild bootstrap procedure for ridge regression. The wild bootstrap ridge regression model is

$$Y^* = X\hat{\beta}(\lambda) + \varepsilon^*,$$

where $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$ with $\varepsilon_i^* = \delta_i \hat{\varepsilon}_i$, where the $\delta_i$s are i.i.d. random variables with $\mathbb{E}[\delta_i] = 0$ and $Var[\delta_i] = 1$ (Mammen, 1993). Similar to least squares regression, the wild bootstrap estimator $\hat{\beta}^*(\lambda)$ of $\hat{\beta}(\lambda)$ is then

$$\begin{aligned}\hat{\beta}^*(\lambda) &= (X'X)^{-1}X'Y^* \\ &= (X'X)^{-1}X'X\hat{\beta}(\lambda) + (X'X)^{-1}X'\varepsilon^* \\ &= \hat{\beta}(\lambda) + (X'X)^{-1}X'\varepsilon^*.\end{aligned}$$

We can use the wild bootstrap estimator to construct a $(1-\alpha)100\%$ confidence region by taking the $(1-\alpha)$ quantile of $(\hat{\beta}^*(\lambda)-\hat{\beta}(\lambda))'X'X(\hat{\beta}^*(\lambda)-\hat{\beta}(\lambda))$. This approach suffers from the same computational challenges as what was faced with least squares regression. It should be noted that, in the penalized regression setting, interpretation of resulting confidence regions and intervals should be carefully considered due to the biased nature of the estimator.

### 3.2.2 Analytic Wild Bootstrap

#### Confidence regions

We implement a similar approach to construct confidence regions for the parameters in ridge regression.

**Theorem 3.2.1.** *For a linear model $Y = X\beta + \varepsilon$ with independent, not necessarily identically distributed errors $\varepsilon$, denote $\hat{\beta}(\lambda)$ as the ridge regression estimator of $\beta$ and the residuals as $\hat{\varepsilon}_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\beta}(\lambda)x_i$. Consider the wild bootstrap model $Y^* = X\hat{\beta}(\lambda) + \varepsilon^*$, where $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$ with $\varepsilon_i^* = \delta_i\hat{\varepsilon}_i$ and the $\delta_i$s are i.i.d. from a symmetric distribution such that $\mathbb{E}[\delta_i] = 0$ and $Var[\delta_i] = 1$, we denote $\hat{\beta}^*(\lambda)$ as the wild bootstrap estimator of $\hat{\beta}(\lambda)$. Furthermore, assume that for $p \geq 2$ and for a constant $B_p$,*

$$\left(\mathbb{E}_\delta \left|\sum_{i=1}^n \delta_i\hat{\varepsilon}_i\right|^p\right)^{1/p} \leq B_p^2 \left(\mathbb{E}_\delta \left|\sum_{i=1}^n \delta_i\hat{\varepsilon}_i\right|^2\right)^{1/2}$$

*for each sequence $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$. Then, for $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,*

$$P\Bigg((\hat{\beta}^*(\lambda) - \hat{\beta}(\lambda))'(X'X + \lambda I)(\hat{\beta}^*(\lambda) - \hat{\beta}(\lambda)) \leq \\ -\log(\alpha^*)\, 4C \left[\sum_{i,j=1}^n h_{ij,\lambda}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}\Bigg) \geq 1 - \alpha,$$

*where $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ is the regularized incomplete beta function, $\theta_1$ and $\theta_2$ are fixed unknown constants, $C$ is a universal constant and $h_{ij,\lambda}$ is the $(i,j)^{th}$ element of $H_\lambda = X(X'X + \lambda I)^{-1}X'$.*

*Proof.* Consider the ridge regression estimator

$$\hat{\beta}(\lambda) = (X'X + \lambda I)^{-1}X'X\hat{\beta},$$

where $\lambda \geq 0$ is the shrinkage parameter, $I$ is the $p \times p$ identity matrix and $\hat{\beta} = (X'X)^{-1}X'Y$. It follows that the wild bootstrap ridge regression estimator is given by

$$\hat{\beta}(\lambda)^* = (X'X + \lambda I)^{-1} X'X \hat{\beta}^*,$$

where $\hat{\beta}^* = \hat{\beta} + (X'X)^{-1}X'\varepsilon^*$. We have that

$$\begin{aligned}
\hat{\beta}(\lambda)^* - \hat{\beta}(\lambda) &= (X'X + \lambda I)^{-1} X'X\hat{\beta}^* - (X'X + \lambda I)^{-1}X'X\hat{\beta} \\
&= (X'X + \lambda I)^{-1} X'X(\hat{\beta} + (X'X)^{-1}X'\varepsilon^* - \hat{\beta}) \\
&= (X'X + \lambda I)^{-1} X'X(X'X)^{-1}X'\varepsilon^* \\
&= (X'X + \lambda I)^{-1} X'\varepsilon^*.
\end{aligned}$$

Thus,

$$\begin{aligned}
(\hat{\beta}(\lambda)^* - \hat{\beta}(\lambda))'(X'X + \lambda I)(\hat{\beta}(\lambda)^* - \hat{\beta}(\lambda)) &= \varepsilon^{*'} X(X'X + \lambda I)^{-1}X'\varepsilon^* \\
&= \varepsilon^{*'} H_\lambda \varepsilon^* \\
&= \sum_{i,j=1}^{n} h_{ij,\lambda}\varepsilon_i^*\varepsilon_j^* \\
&= \sum_{i,j=1}^{n} h_{ij,\lambda}\delta_i\delta_j\hat{\varepsilon}_i\hat{\varepsilon}_j \\
&= T^2.
\end{aligned}$$

The proof follows the same procedure as seen in the proof of Theorem 2.5.1.

□

**Linear contrasts**

Similarly, as in Section 2.5.2, in order to consider the linear contrast $c'(\hat{\beta}^*(\lambda) - \hat{\beta}(\lambda))$, we introduce the following theorem.

**Theorem 3.2.2.** *Under the same settings as Theorem 3.2.1, let $c \in \mathbb{R}^p$ be a fixed contrast vector and $v_\lambda = (v_{1,\lambda}, \ldots, v_{n,\lambda})' = c'(X'X + \lambda I)^{-1}X'$. Furthermore, assume that for a constant $B_p$,*

$$\left[\mathbb{E}_\delta \left|\sum_{i=1}^{n} v_{i,\lambda}\delta_i\hat{\varepsilon}_i\right|^p\right]^{1/p} \leq B_p^2 \left[\sum_{i=1}^{n} v_{i,\lambda}^2\hat{\varepsilon}_i^2\right]^{1/2}$$

*for each sequence* $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$*. Then, for* $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$*,*

$$P\left(\left|c'(\hat{\beta}^*(\lambda) - \hat{\beta}(\lambda))\right| \leq \left[-\log(\alpha^*)\, 4\left[\sum_{i=1}^{n} v_{i,\lambda}^2 \hat{\varepsilon}_i^2\right]^{1/2}\right]^{1/2}\right) \geq 1 - \alpha,$$

*where* $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$*,* $I(\alpha; \theta_1, \theta_2)$ *is the regularized incomplete beta function and* $\theta_1$*,* $\theta_2$ *are fixed unknown constants.*

*Proof.* Consider the linear contrast

$$\begin{aligned} c'(\hat{\beta}^*(\lambda) - \hat{\beta}(\lambda)) &= c'(X'X + \lambda I)^{-1} X' \varepsilon^* \\ &= v_\lambda' \varepsilon^* \\ &= \sum_{i=1}^{n} v_{i,\lambda} \varepsilon_i^* \\ &= \sum_{i=1}^{n} v_{i,\lambda} \delta_i \hat{\varepsilon}_i \\ &= T, \end{aligned}$$

where $v_\lambda = (v_{1,\lambda}, \ldots, v_{n,\lambda})' = c'(X'X + \lambda I)^{-1} X'$. The proof follows the same procedure as seen in the proof of Theorem 2.5.2. □

### 3.2.3 Simulations

We simulated data for ridge regression under the same settings outlined in Section 2.6.1 using the `glmnet` package to select the optimal value for $\lambda$ via cross-validation. As can be seen in Figure 3.1 and Table 3.1, the performance of the methodologies is similar to that of least squares regression.

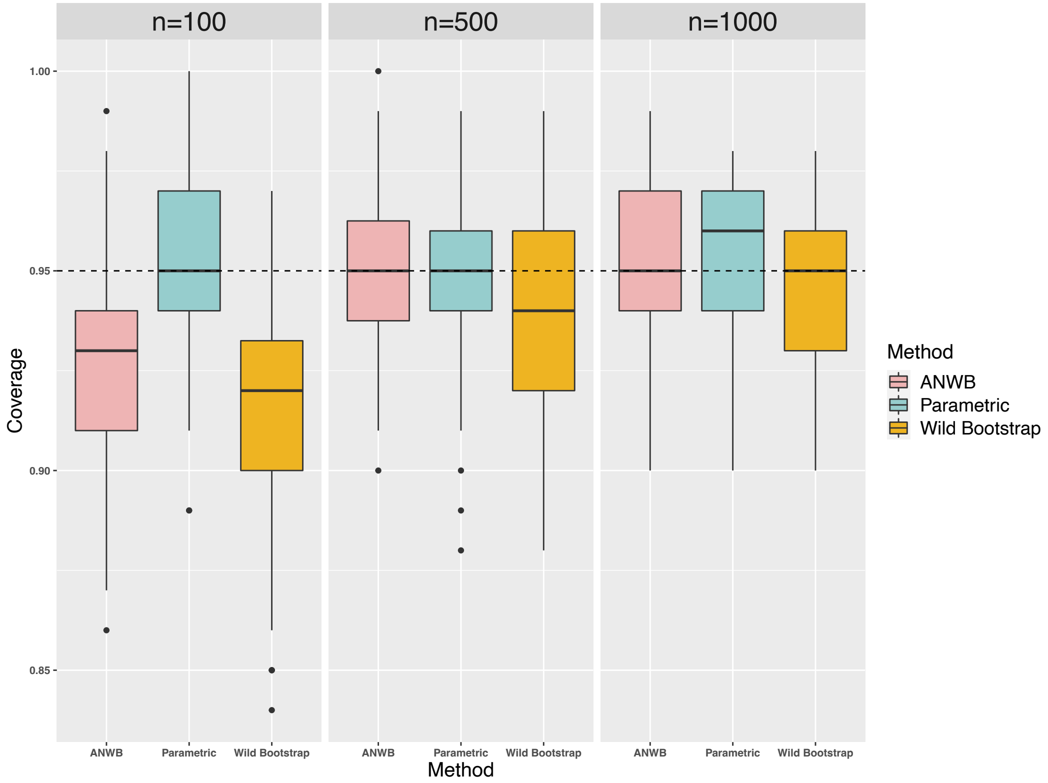

**Figure 3.1:** Coverage of 95% ANWB, parametric and wild bootstrap confidence regions for homoscedastic data

**Table 3.1:** Mean and standard deviation of the coverages of 95% ANWB, parametric and wild bootstrap confidence regions for homoscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| ANWB | 0.929 (0.023) | 0.95 (0.022) | 0.953 (0.019) |
| Parametric | 0.952 (0.022) | 0.95 (0.022) | 0.953 (0.019) |
| Wild Bootstrap | 0.915 (0.027) | 0.942 (0.022) | 0.949 (0.019) |

### 3.2.4 Discussion

As can be seen from the simulations, bootstrapping ridge regression estimators may provide good results in certain situations, but it is a somewhat naive approach as it ignores a fundamental assumption of bootstrapping in general. In ridge regression, the estimator is biased which is problematic as it lacks a pivot (Vinod, 1995). In least squares regression, $(\hat{\beta} - \beta)$ is a pivotal quantity, meaning that its sampling distribution does not depend on $\beta$, so no issues arise. However, in ridge regression, $(\hat{\beta}(\lambda) - \beta)$ is not a pivotal quantity so it may be dangerous to apply bootstrapping in this scenario. Vinod (1987) notes that bootstrapping may be a useful tool for providing information about a sampling distribution if we know that the ridge regression estimator, $\hat{\beta}(\lambda)$, is estimating $\beta$ well (i.e., with little bias). Nevertheless, this is clearly not always the case so caution should be taken when using this naive approach as it may not always provide accurate results. In the next section, we introduce Beran (1987)'s double bootstrap as a means to overcome the lack of pivot problem in ridge regression.

## 3.3 The Double Bootstrap

The double bootstrap was developed by Beran (1987) and can be used to help address the lack of pivot problem that researchers face when dealing with the estimator in ridge regression. Essentially, by doing two layers of bootstrapping, we can improve the convergence of the single bootstrap when working with a biased estimator, such as in ridge regression (McCullough and Vinod, 1998). Vinod (1995) discusses the application of the double bootstrap in ridge regression utilizing the residual bootstrap. Without loss of generality, we adopt two stages of the wild bootstrap as a way to perturb the residuals instead of the classic residual bootstrap. Applying the double bootstrap gets us a higher order of accuracy for respective confidence regions (McCullough and Vinod, 1998).

### 3.3.1 Double wild bootstrap

Below is the outline of the double wild bootstrap model for ridge regression. Note that, going forward, when we use the term double bootstrap we are referring to the double wild bootstrap.

**Algorithm 2** The Double Bootstrap for Ridge Regression

```
First-stage:
    For $j = 1, 2, \ldots, J$ compute a vector of perturbed residuals $\varepsilon^*_j$
    Calculate perturbed responses $Y^*_j = X\hat{\beta}(\lambda) + \varepsilon^*_j$.
    Estimate regression coefficients
        $\hat{\beta}(\lambda)^*_j = (X'X + \lambda I)^{-1}X'Y^*_j$.

Second-stage:
    For each first-stage bootstrap sample $j$, for $k = 1, 2, \ldots, K$ compute
    $\varepsilon^{**}_{jk}$, where $\hat{\varepsilon}^{**}_{jk} = \varepsilon^*_j\delta$ and $\delta$ is an $n$-vector of Rademacher realizations
    Recalculate responses $Y^{**}_{jk} = X\hat{\beta}^*_j(\lambda) + \varepsilon^{**}_{jk}$
    Re-estimate coefficients as
        $\hat{\beta}(\lambda)^{**}_{jk} = (X'X + \lambda I)^{-1}X'Y^{**}_{jk}$.

For each first-stage bootstrap sample $j$, compute the median of $Q$,
where $Q_j$ is the $(1 - \alpha)$ quantile of
        $(\hat{\beta}(\lambda)^{**}_{jk} - \hat{\beta}(\lambda)^*_j)'(X'X + \lambda I)(\hat{\beta}(\lambda)^{**}_{jk} - \hat{\beta}(\lambda)^*_j)$
```

We will now look at a small example to further motivate the use of the double bootstrap in ridge regression. We simulated our data from the following model

$$Y = X\beta + \varepsilon,$$

where $n = 30$, $p = 3$, $\varepsilon \sim \mathcal{N}(0, 3)$, $X$ was generated from Uniform[0,5] and $\beta = (2, 0, -1.5)'$. Here, $\lambda = 1.17$ was selected via cross-validation using the `glmnet` package. For the wild bootstrap, we use $B = 1000$ bootstrap replications and construct 95% confidence intervals for $\beta_1 = 2$, comparing the performance of the single and double wild bootstrap over $r = 100$ replications.

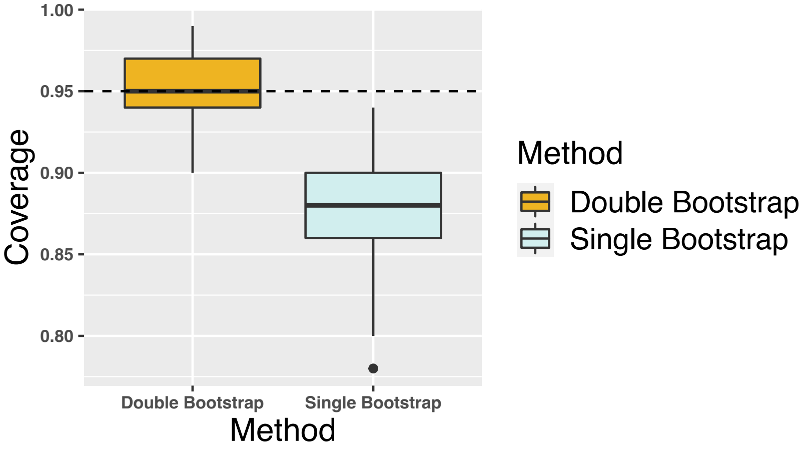

**Figure 3.2:** Coverage of 95% confidence intervals for the single and double bootstrap

**Table 3.2:** Mean and standard deviation of the coverages of 95% confidence intervals for the single and double bootstrap

| Method | Mean | Standard Deviation |
|---|---|---|
| Double Bootstrap | 0.953 | 0.021 |
| Single Bootstrap | 0.876 | 0.029 |

We can see that, due to the biased nature of the estimator, the single bootstrap is much too liberal in terms of coverage with a mean coverage of 87.6%, grossly underestimating the desired 95% coverage. The implementation of the double bootstrap drastically improves the accuracy of the coverage with a mean of approximately 95.3%. Although there is a chance the single bootstrap could produce the desired coverage, this is a naive approach and the double bootstrap confidence intervals better capture the desired confidence level in this setting.

Much like what was done in least squares regression, we are now able to utilize the double wild bootstrap estimator to construct more accurate $(1 - \alpha)100\%$

confidence regions by taking the $(1-\alpha)$ quantile of $(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*})'(X'X + \lambda I)(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*})$.

### 3.3.2 Fast double bootstrap

Although the double bootstrap helps us address the lack of pivot problem in ridge regression, performing two layers of bootstrapping comes at a dramatic computational cost. For $B$ first-stage bootstrap replications and $B'$ second-stage bootstrap replications, the double bootstrap requires $B'' = BB'$ total replications, which is quite the computational demand, especially as the dimensions of the data increase.

In response to this problem, the fast double bootstrap was developed (Davidson and MacKinnon, 2007). Computationally, the fast double bootstrap is practically equivalent to the single bootstrap, so it is much more appealing from that standpoint (Davidson and MacKinnon, 2007). Only one second-stage bootstrap replication is required for the fast double bootstrap which is why it requires much less time to implement.

The gain in efficiency comes with an added assumption of independence. The fast double bootstrap requires the assumption that, given a test statistic of interest $\tau$, the distribution of $\tau_{jk}^{**}$ is independent of $\tau_j^{*}$ (Davidson & MacKinnon, 2007). In our setting, this is equivalent to the assumption that the distribution of $\hat{\beta}(\lambda)_{jk}^{**}$ is independent of $\hat{\beta}(\lambda)_j^{*}$. Given the data, this assumption holds as we are simply considering a linear transformation of independent Rademacher random variables. We outline the fast double bootstrap for ridge regression below in Algorithm 3. Going forward, when mentioning the double bootstrap it is implied that we are referring to the fast double bootstrap as it is much more computationally efficient and exhibits analogous performance to the double bootstrap.

**Algorithm 3** The Fast Double Bootstrap for Ridge Regression

```
First-stage:
    For $j = 1, 2, \dots, J$ compute a vector of perturbed residuals $\varepsilon_j^*$
    Calculate perturbed responses $Y_j^* = X\hat{\beta}(\lambda) + \varepsilon_j^*$.
    Estimate regression coefficients
        $\hat{\beta}(\lambda)_j^* = (X'X + \lambda I)^{-1} X' Y_j^*$.

Second-stage:
    For each first-stage bootstrap sample $j$, compute
    $\varepsilon_j^{**}$, where $\varepsilon_j^{**} = \varepsilon_j^* \delta$ and $\delta$ is an $n$-vector of Rademacher realizations
    Recalculate responses $Y_j^{**} = X\hat{\beta}_j^*(\lambda) + \varepsilon_j^{**}$
    Re-estimate coefficients as
        $\hat{\beta}(\lambda)_j^{**} = (X'X + \lambda I)^{-1} X' Y_j^{**}$.

Compute the $(1 - \alpha)$ quantile of $Q$, where
        $Q_j = (\hat{\beta}(\lambda)_j^{**} - \hat{\beta}(\lambda)_j^*)'(X'X + \lambda I)(\hat{\beta}(\lambda)_j^{**} - \hat{\beta}(\lambda)_j^*)$
```

## 3.4 Analytic Fast Double Bootstrap

### 3.4.1 Confidence regions

We now adapt our approach discussed in Section 2.5 to construct confidence regions for the parameters in ridge regression using the analytic fast double bootstrap (AFDB).

**Theorem 3.4.1.** *For a linear model $Y = X\beta + \varepsilon$ with independent, not necessarily identically distributed errors $\varepsilon$, denote $\hat{\beta}(\lambda)$ as the ridge regression estimator of $\beta$ and the residuals as $\hat{\varepsilon}_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\beta}(\lambda)x_i$. Consider the wild bootstrap model $Y^* = X\hat{\beta}(\lambda) + \varepsilon^*$, where $\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)'$ with $\varepsilon_i^* = \delta_{i,1}\hat{\varepsilon}_i$ and the $\delta_{i,1}$s are i.i.d. from a symmetric distribution such that $\mathbb{E}[\delta_{i,1}] = 0$ and $Var[\delta_{i,1}] = 1$, we denote $\hat{\beta}^*(\lambda)$ as the wild bootstrap estimator of $\hat{\beta}(\lambda)$. Also, consider the double wild bootstrap model $Y^{**} = X\hat{\beta}(\lambda)^* + \varepsilon^{**}$,* *where $\varepsilon^{**} = (\varepsilon_1^{**}, \dots, \varepsilon_n^{**})'$ with $\varepsilon_i^{**} = \delta_{i,2}\varepsilon_i^*$ and the $\delta_{i,2}$s are i.i.d. from a symmetric distribution such that $\mathbb{E}[\delta_{i,2}] = 0$ and $Var[\delta_{i,2}] = 1$, we denote $\hat{\beta}(\lambda)^{**}$ as the double wild bootstrap estimator of $\hat{\beta}(\lambda)^*$.Furthermore, assume that for $p \geq 2$ and for a constant $B_p$,*

$$\left(\mathbb{E}_\delta \left|\sum_{i=1}^{n} \delta_{i,2}\varepsilon_i^*\right|^p\right)^{1/p} \leq B_p^2 \left(\mathbb{E}_\delta \left|\sum_{i=1}^{n} \delta_{i,2}\varepsilon_i^*\right|^2\right)^{1/2}$$

*for each sequence $\hat{\varepsilon}_1^*, \dots, \hat{\varepsilon}_n^* \in \mathbb{R}$. Then, for $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,*

$$P\Bigg((\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^*)'(X'X + \lambda I)(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^*) \leq \\ -\log(\alpha^*)\, 4C \left[\sum_{i,j=1}^{n} h_{ij,\lambda}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}\Bigg) \geq 1 - \alpha,$$

*where $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ is the regularized incomplete beta function, $\theta_1$ and $\theta_2$ are fixed unknown constants, $C$ is a universal constant and $h_{ij,\lambda}$ is the $(i,j)^{th}$ element of $H_\lambda = X(X'X + \lambda I)^{-1}X'$.*

*Proof.* Consider the ridge regression estimator

$$\hat{\beta}(\lambda) = (X'X + \lambda I)^{-1}X'X\hat{\beta},$$

where $\lambda \geq 0$ is the shrinkage parameter, $I$ is the $p \times p$ identity matrix and $\hat{\beta} = (X'X)^{-1}X'Y$. It follows that the wild bootstrap ridge regression estimator is given by

$$\hat{\beta}(\lambda)^* = (X'X + \lambda I)^{-1}X'X\hat{\beta}^*,$$

where $\hat{\beta}^* = \hat{\beta} + (X'X)^{-1}X'\varepsilon^*$. Similarly, the double wild bootstrap ridge regression estimator is given by

$$\hat{\beta}(\lambda)^{**} = (X'X + \lambda I)^{-1}X'X\hat{\beta}^{**},$$

where $\hat{\beta}^{**} = \hat{\beta}^{*} + (X'X)^{-1}X'\varepsilon^{**}$.We have that

$$\begin{aligned}
\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*} &= (X'X+\lambda I)^{-1}X'X\hat{\beta}^{**} - (X'X+\lambda I)^{-1}X'X\hat{\beta}^{*} \\
&= (X'X+\lambda I)^{-1}X'X(\hat{\beta}^{*} + (X'X)^{-1}X'\varepsilon^{**} - \hat{\beta}^{*}) \\
&= (X'X+\lambda I)^{-1}X'X(X'X)^{-1}X'\varepsilon^{**} \\
&= (X'X+\lambda I)^{-1}X'\varepsilon^{**}.
\end{aligned}$$

Thus,

$$\begin{aligned}
(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*})'(X'X+\lambda I)(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*}) &= \varepsilon^{**'}X(X'X+\lambda I)^{-1}X'\varepsilon^{**} \\
&= \varepsilon^{**'}H_{\lambda}\varepsilon^{**} \\
&= \sum_{i,j=1}^{n} h_{ij,\lambda}\varepsilon_i^{**}\varepsilon_j^{**} \\
&= \sum_{i,j=1}^{n} h_{ij,\lambda}\delta_{i,1}\delta_{j,1}\delta_{i,2}\delta_{j,2}\hat{\varepsilon}_i\hat{\varepsilon}_j \\
&= T^2.
\end{aligned}$$

The proof follows the same procedure as seen in the proof of Theorem 2.5.1.

□

### 3.4.2 Linear contrasts

As was introduced in Chapter 1, we can similarly consider the linear contrast $c'(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*})$ by introducing the following theorem.

**Theorem 3.4.2.** *Under the same settings as Theorem 3.4.1, let $c \in \mathbb{R}^p$ be a fixed contrast vector and $v_\lambda = (v_{1,\lambda}, \ldots, v_{n,\lambda})' = c'(X'X+\lambda I)^{-1}X'$. Furthermore, assume that for a constant $B_p$,*

$$\left[\mathbb{E}_{\delta}\left|\sum_{i=1}^{n} v_{i,\lambda}\delta_{i,2}\varepsilon_i^{*}\right|^{p}\right]^{1/p} \leq B_p^2\left[\sum_{i=1}^{n} v_{i,\lambda}^2\hat{\varepsilon}_i^2\right]^{1/2}$$

*for each sequence* $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$. *Then, for* $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,

$$P\left(\left|c'(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*})\right| \leq \left[-\log(\alpha^*)\, 4\left[\sum_{i=1}^{n} v_{i,\lambda}^2 \hat{\varepsilon}_i^2\right]^{1/2}\right]^{1/2}\right) \geq 1-\alpha,$$

*where* $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ *is the regularized incomplete beta function and* $\theta_1$, $\theta_2$ *are fixed unknown constants.*

*Proof.* Consider the linear contrast

$$\begin{aligned} c'(\hat{\beta}(\lambda)^{**} - \hat{\beta}(\lambda)^{*}) &= c'(X'X + \lambda I)^{-1} X' \varepsilon^{**} \\ &= v_\lambda' \varepsilon^{**} \\ &= \sum_{i=1}^{n} v_{i,\lambda} \varepsilon_i^{**} \\ &= \sum_{i=1}^{n} v_{i,\lambda} \delta_{i,2} \varepsilon_i^{*} \\ &= T, \end{aligned}$$

where $v_\lambda = (v_{1,\lambda}, \ldots, v_{n,\lambda})' = c'(X'X + \lambda I)^{-1} X'$. The proof follows the same procedure as seen in the proof of Theorem 2.5.2. □

## 3.5 Simulations

We simulated our data from the following linear model

$$Y = X\beta + \varepsilon,$$

where $n \in \{100, 500, 1000\}$, $p = 5$, $\varepsilon \sim \mathcal{N}(0, 3)$, $X$ is generated from Uniform[0,5] and $\beta$ is generated from Uniform[0,3]. We take the $\delta_i$s to be i.i.d. Rademacher random variables with $P(\delta_i = 1) = P(\delta_i = -1) = 1/2$. We also consider heteroscedastic data simulated with $\varepsilon \sim \mathcal{N}(0, \sigma_i^2)$, where $\sigma_i^2 = x_{i1}^2$ (Flachaire, 2005). Simulation results are presented for $\alpha = 0.05$.

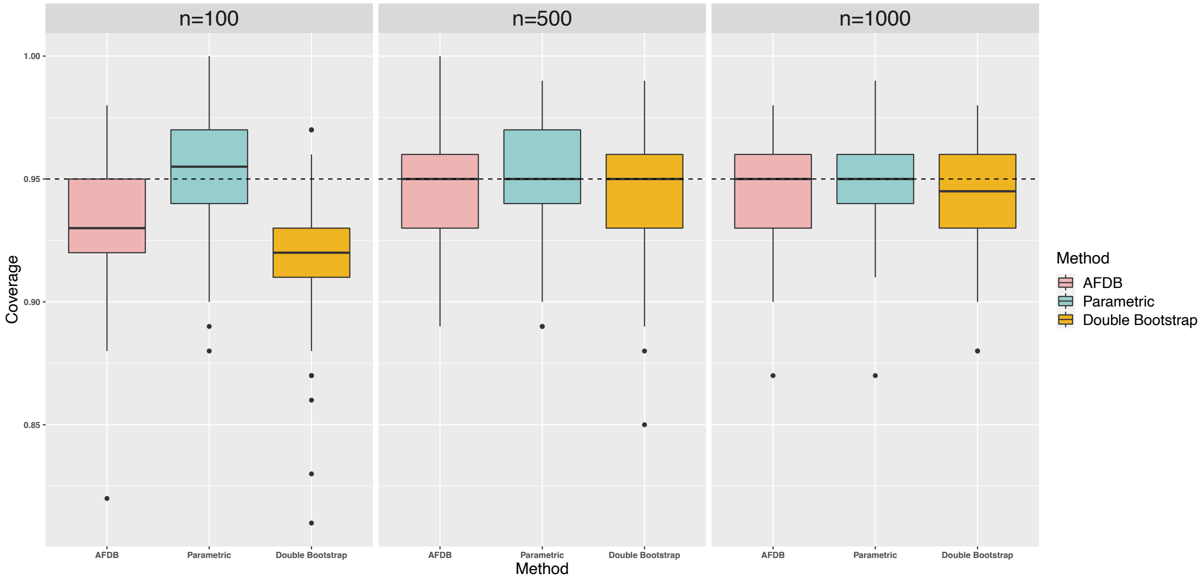

**Figure 3.3:** Coverage of 95% AFDB, parametric and double bootstrap confidence regions for homoscedastic data

**Table 3.3:** Mean and standard deviation of the coverages of 95% AFDB, parametric and double bootstrap confidence regions for homoscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| AFDB | 0.931 (0.027) | 0.948 (0.023) | 0.946 (0.021) |
| Parametric | 0.952 (0.022) | 0.951 (0.024) | 0.948 (0.02) |
| Double Bootstrap | 0.917 (0.026) | 0.942 (0.027) | 0.944 (0.021) |

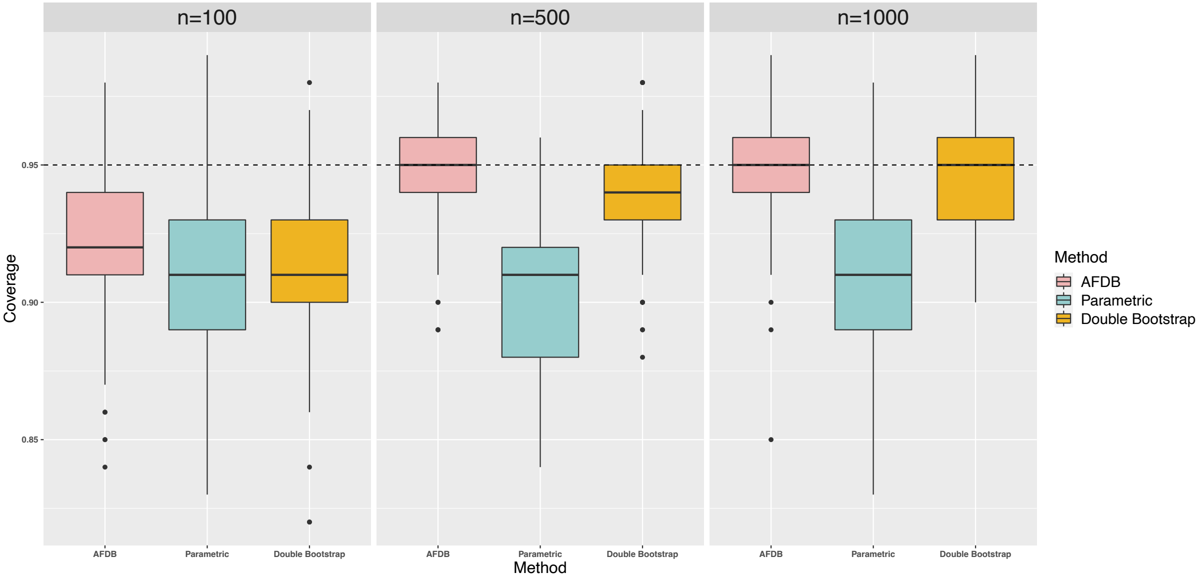

**Figure 3.4:** Coverage of 95% AFDB, parametric and double bootstrap confidence regions for heteroscedastic data

**Table 3.4:** Mean and standard deviation of the coverages of 95% AFDB, parametric and double bootstrap confidence regions for heteroscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| AFDB | 0.921 (0.028) | 0.946 (0.021) | 0.949 (0.022) |
| Parametric | 0.910 (0.030) | 0.905 (0.028) | 0.908 (0.029) |
| Double Bootstrap | 0.910 (0.029) | 0.941 (0.020) | 0.948 (0.021) |

## 3.6 Data Example

In this section, we will demonstrate our approach on a real data set focusing on energy efficiency of residential buildings. In particular, this data set investigates the relationship between the shape of a building and its heating load, the response variable (Tsanas and Xifara, 2012). This data set has 768 observations and 8 predictor variables: relative compactness ($X_1$), surface area ($X_2$), wall area ($X_3$), roof area ($X_4$), overall height ($X_5$), orientation ($X_6$), glazing area ($X_7$) and glazing area distribution ($X_8$). One can imagine that these variables might be high correlated, leading to the issue of multicollinearity when fitting a linear regression model. For example, relative compactness and surface area are strongly negatively correlated with $r = -0.991$, as when surface area increases, the compactness of the building decreases.

A model was fit to the data using ridge regression with tuning parameter $\lambda = 0.897$ selected via cross-validation. After model selection, $X_4$ and $X_6$ were removed resulting in an estimated coefficient vector of $\hat{\beta}(\lambda) = (0.70, -1.91, 3.22, 5.13, 2.32, 0.34)'$. Below, 95% AFDB, parametric and double bootstrap confidence regions are plotted for regression coefficients $\beta_3$ and $\beta_4$ in Figure 3.5; all three methods produce similar confidence regions.

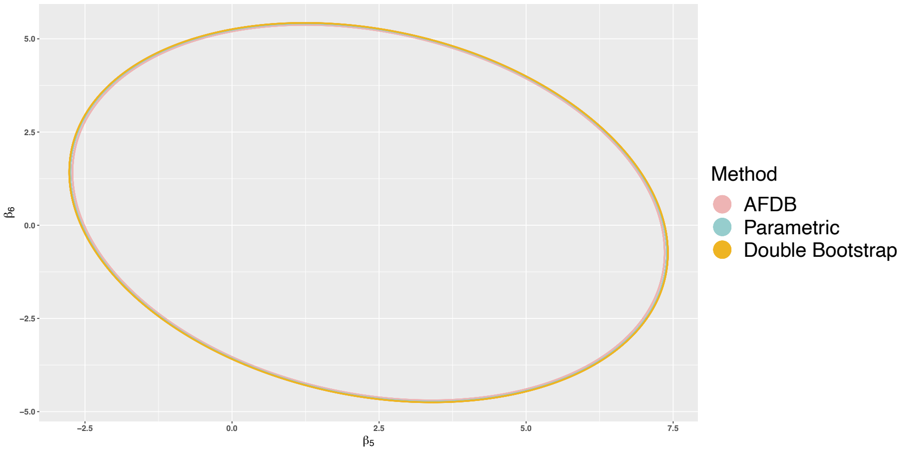

**Figure 3.5:** 95% confidence regions for energy data set

## 3.7 Discussion

This chapter focused on bootstrapping in the ridge regression setting. Initially, the single wild bootstrap was discussed in this context, and while it may perform well in certain parameter settings, it is not prudent to use it for the biased parameters in ridge regression. As a solution, the double bootstrap was proposed as a better estimate of the sampling distribution for ridge coefficients. However, doing multiple rounds of bootstrapping is highly impractical, especially with high-dimensional data. Thus, the fast double bootstrap was proposed, with computational run times on par with the single bootstrap.

While the fast double bootstrap is an improvement over the double bootstrap in terms of computational time, it is still relatively inefficient. The analytic fast double bootstrap (AFDB) was introduced in the context of ridge regression. Through simulation studies and a real data example, we were able to establish the similarities in performance between the AFDB and the fast double bootstrap, noting the benefits of the AFDB from a computational standpoint.

# Chapter 4

# Analytic Bootstrap for LASSO Regression

## 4.1 Introduction

The LASSO regression estimator $\hat{\beta}^{lasso}$ is a shrinkage estimator with an $L_1$ penalty such that, as defined in Efron and Hastie (2016),

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ ||Y - X\beta||_2^2 + \lambda \, ||\beta||_1 \right\}.$$

However, unlike ridge regression, there is no closed form expression for LASSO coefficients. Thus, extending the approach of the analytic wild bootstrap to LASSO is not as straight-forward.

This chapter considers the implementation of the bootstrap, both computationally and analytically, in LASSO regression. Section 4.2 discusses issues that arise when bootstrapping in LASSO estimators and introduces the concept of approximate LASSO solutions as a work around to apply to double bootstrap in this setting. In Section 4.3, we highlight the advantages of using an iterative scheme rather than the ridge approximation alone to estimate coefficients. Section 4.4 describes the AFDB in the context of LASSO regression as a means to compute confidence regions. Sections 4.5 and 4.6 demonstrate the performance of various approaches to constructing confidence regions and intervals in LASSO regression. Finally, Section 4.7 wraps up with a summary and discussion of the findings from this chapter.

## 4.2 Wild bootstrap

We have shown that the wild bootstrap works well in least squares regression. Furthermore, in LASSO regression, Chatterjee and Lahiri (2010) show that in the presence of zero coefficients, the bootstrap may fail to be consistent. As a result, they developed a modified bootstrap that, under some assumptions, is consistent. Alternatively, much work has been done exploring the debiased LASSO estimator. First introduced by Zhang and Zhang (2014), the debiased LASSO estimator has become a popular way to conduct statistical inference in high-dimensional models (Van De Geer, 2019; Javanmard and Montanari, 2018; Li, 2020).

Current approaches for conducting statistical inference in penalized regression may be complicated to implement as well as computationally expensive. In this chapter, we adapt the idea of the AFDB to LASSO regression using approximate LASSO solutions. Sartori (2011) outlines an approach for approximating LASSO solutions which we describe in detail below. First, Tibshirani (1996) used the fact that we can rewrite the LASSO penalty as

$$\sum_{j=1}^{p} |\beta_j| = \sum_{j=1}^{p} \beta_j^2/|\beta_j|.$$

This enables us to express the LASSO estimator as

$$\begin{aligned}
\hat{\beta}^{lasso} &= \underset{\beta}{\operatorname{argmin}} \left\{ ||Y - X\beta||^2 + \lambda \, ||\beta|| \right\} \\
&= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2/|\beta_j| \\
&= \underset{\beta}{\operatorname{argmin}} (Y - X\beta)'(Y - X\beta) + \lambda \beta' \Lambda_\beta \beta,
\end{aligned}$$

where $\Lambda_\beta = diag(1/|\beta_1|, \ldots, 1/|\beta_p|)$. However, in LASSO regression one or more of the coefficients $\beta_1, \ldots, \beta_p$ may be zero. To circumvent the problem of dividing by zero, we will redefine $\Lambda_\beta = diag(|\beta_1|, \ldots, |\beta_p|)$ and introduce $\Lambda_\beta^-$: the generalized inverse of $\Lambda_\beta$. That is, for $j = 1, \ldots, p$,

$$\Lambda_\beta^- = \begin{cases} 1/|\beta_j|, & |\beta_j| > 0 \\ 0, & |\beta_j| = 0. \end{cases}$$

Thus, our LASSO estimator becomes

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}}(Y - X\beta)'(Y - X\beta) + \lambda\beta'\Lambda_\beta^-\beta.$$

Consequently, this can be approximated by ridge regression as

$$\hat{\beta}^{lasso} \approx (X'X + \lambda\Lambda_{\hat{\beta}}^-)^{-1}X'Y,$$

where $\Lambda_{\hat{\beta}} = diag(|\hat{\beta}_1|, \ldots, |\hat{\beta}_p|)$ and $\hat{\beta}^{lasso} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$. It should be noted that, using this ridge approximation, regression coefficients cannot be set exactly to zero. Nevertheless, this approach enables us to apply the double wild bootstrap as a means to construct confidence regions. As was outlined in Section 3.2, the single bootstrap may not yield the best results due to the bias of the ridge regression estimator. Therefore, we continue to adopt the double bootstrap as a means to estimate the sampling distribution. Thus, we are now able to utilize the double bootstrap to construct a $(1 - \alpha)100\%$ confidence region by taking the $(1 - \alpha)$ quantile of

$$(\hat{\beta}^{lasso^{**}} - \hat{\beta}^{lasso^{*}})'(X'X + \lambda\Lambda_{\hat{\beta}}^-)(\hat{\beta}^{lasso^{**}} - \hat{\beta}^{lasso^{*}}).$$

## 4.3 Iterative Wild Bootstrap for LASSO Regression

In general, the ridge approximation works well for estimating LASSO coefficients. However, the approximation can be improved for the case of the zero coefficients (Sartori, 2011). This can be achieved by considering an iterative ridge regression algorithm for approximating the LASSO coefficients (Tibshirani, 1996; Sartori, 2011). Denoted as the ridge-IWLS approximation, we are able to obtain improved approximations for LASSO coefficients, especially those coefficients that would be otherwise set to zero in LASSO regression. Described below is the algorithm as seen in (Sartori, 2011).

**Algorithm 4** The Ridge-IWLS Approximation

```
    Calculate adjusted responses $z_i^{(k)} = x_i\hat{\beta}^{(k)} + y_i - \hat{y}_i^{(k)}$.

    Denote $W^{(k)} = \text{diag}(1, \ldots, 1)$ and $\Lambda_{\hat{\beta}}^{(k)} = \text{diag}\left(|\hat{\beta}_1^{(k)}|, \ldots, |\hat{\beta}_1^{(k)}|\right)$

    Update coefficients
        $\hat{\beta}^{(k+1)} = \left(X'W^{(k)}X + \lambda\Lambda_{\hat{\beta}}^{(k)-}\right)^{-1} X'W^{(k)}z^{(k)}$.

    Repeat until $|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}| < \varepsilon$ for $\varepsilon = 0.0001$.
```

To further motivate the use of the iterative approach, consider the following simulated example with $\beta = (3, 0, -2, 0.5, 2.5)'$ and $n = 1000$. From table 4.1, although both approaches give good approximations to the estimated coefficients from LASSO, in the case of a zero coefficient, the ridge-IWLS clearly gives a much better approximation than just the ridge approximation alone.

**Table 4.1:** Comparing coefficient estimates from LASSO, ridge approximation and ridge-IWLS with the true $\beta$

| $\beta$ | $\hat{\beta}^{lasso}$ | $\hat{\beta}$ (ridge approx.) | $\hat{\beta}$ (ridge-IWLS) |
|---|---|---|---|
| 3 | 2.598604 | 2.598882 | 2.598604 |
| 0 | 0.000000 | $-0.038951$ | $-0.000001$ |
| $-2$ | $-1.954168$ | $-1.956218$ | $-1.954169$ |
| 0.5 | 0.000000 | 0.075460 | 0.000127 |
| 2.5 | 2.954878 | 2.951977 | 2.954874 |

## 4.4 Analytic Wild Bootstrap for LASSO Regression

In this section, we consider the construction of analytic confidence regions using the AFDB in an similar manner to what was done for the parameters in ridge regression in Chapter 3.

**Theorem 4.4.1.** *For a linear model $Y = X\beta + \varepsilon$ with independent, not necessarily identically distributed errors $\varepsilon$, denote $\hat{\beta}^{lasso}$ as the LASSO estimator of $\beta$ and the residuals as $\hat{\varepsilon}_i = y_i - \hat{y}_i$, where $\hat{y}_i = \hat{\beta}^{lasso} x_i$. Consider the wild bootstrap model $Y^* = X\hat{\beta}^{lasso} + \varepsilon^*$, where $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)'$ with $\varepsilon_i^* = \delta_i \hat{\varepsilon}_i$ and the $\delta_i$s are i.i.d. from a symmetric distribution such that $\mathbb{E}[\delta_i] = 0$ and $Var[\delta_i] = 1$, we denote $\hat{\beta}^{lasso^*}$ as the wild bootstrap estimator of $\hat{\beta}^{lasso}$. Also, consider the double wild bootstrap model $Y^{**} = X\hat{\beta}^{lasso^*} + \varepsilon^{**}$, where $\varepsilon^{**} = (\varepsilon_1^{**}, \ldots, \varepsilon_n^{**})'$ with $\varepsilon_i^{**} = \delta_{i,2}\varepsilon_i^*$ and the $\delta_{i,2}$s are i.i.d. from a symmetric distribution such that $\mathbb{E}[\delta_{i,2}] = 0$ and $Var[\delta_{i,2}] = 1$, we denote $\hat{\beta}^{lasso^{**}}$ as the double wild bootstrap estimator of $\hat{\beta}^{lasso^*}$. Furthermore, assume that for $p \geq 2$ and for a constant $B_p$,*

$$\left(\mathbb{E}_\delta \left|\sum_{i=1}^{n} \delta_{i,2}\varepsilon_i^*\right|^p\right)^{1/p} \leq B_p^2 \left(\mathbb{E}_\delta \left|\sum_{i=1}^{n} \delta_{i,2}\varepsilon_i^*\right|^2\right)^{1/2}$$

*for each sequence $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$. Then, for $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,*

$$P\left((\hat{\beta}^{lasso^{**}} - \hat{\beta}^{lasso^*})'(X'X + \lambda\Lambda_\beta^-)(\hat{\beta}^{lasso^{**}} - \hat{\beta}^{lasso^*}) \leq -\log(\alpha^*)\, 4C \left[\sum_{i,j=1}^{n} h_{ij,lasso}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2\right]^{1/2}\right) \geq 1 - \alpha,$$

*where $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ is the regularized incomplete beta function, $\theta_1$ and $\theta_2$ are fixed unknown constants, $C$ is a universal constant and $h_{ij,lasso}$ is the $(i,j)^{th}$ element of $H_\lambda = X(X'X + \lambda\Lambda_\beta^-)^{-1}X'$.*

*Proof.* The proof follows the same procedure as seen in the proof of Theorem 3.4.1. □

**Theorem 4.4.2.** *Under the same settings as Theorem 4.4.1, let $c \in \mathbb{R}^p$ be a fixed contrast vector and $v_\lambda = (v_{1,\lambda}, \ldots, v_{n,\lambda})' = c'(X'X + \lambda I)^{-1}X'$. Furthermore, assume that for a constant $B_p$,*

$$\left[\mathbb{E}_\delta \left|\sum_{i=1}^{n} v_{i,\lambda}\delta_{i,2}\varepsilon_i^*\right|^p\right]^{1/p} \leq B_p^2 \left[\sum_{i=1}^{n} v_{i,\lambda}^2 \hat{\varepsilon}_i^2\right]^{1/2}$$

*for each sequence $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$. Then, for $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,*

$$P\left(\left|c'(\hat{\beta}^{lasso^{**}} - \hat{\beta}^{lasso^*})\right| \leq \left[-\log\left(\alpha^*\right) 4 \left[\sum_{i=1}^{n} v_{i,\lambda}^2 \hat{\varepsilon}_i^2\right]^{1/2}\right]^{1/2}\right) \geq 1 - \alpha,$$

*where $\alpha^* = I\left(\alpha \left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ is the regularized incomplete beta function and $\theta_1$, $\theta_2$ are fixed unknown constants.*

*Proof.* The proof follows the same procedure as seen in the proof of Theorem 3.4.2. □

## 4.5 Simulations

We simulated our data from the following linear model

$$Y = X\beta + \varepsilon,$$

where $n \in \{100, 500, 1000\}$, $p = 5$, $\varepsilon \sim \mathcal{N}(0, 3)$, $X$ is generated from Uniform[0,5] and $\beta$ is generated from Uniform[0,3]. We take the $\delta_i$s to be i.i.d. Rademacher random variables with $P(\delta_i = 1) = P(\delta_i = -1) = 1/2$. We also consider heteroscedastic data simulated with $\varepsilon \sim \mathcal{N}(0, \sigma_i^2)$, where $\sigma_i^2 = x_{i1}^2$ (Flachaire, 2005). Simulation results are presented for $\alpha = 0.05$.

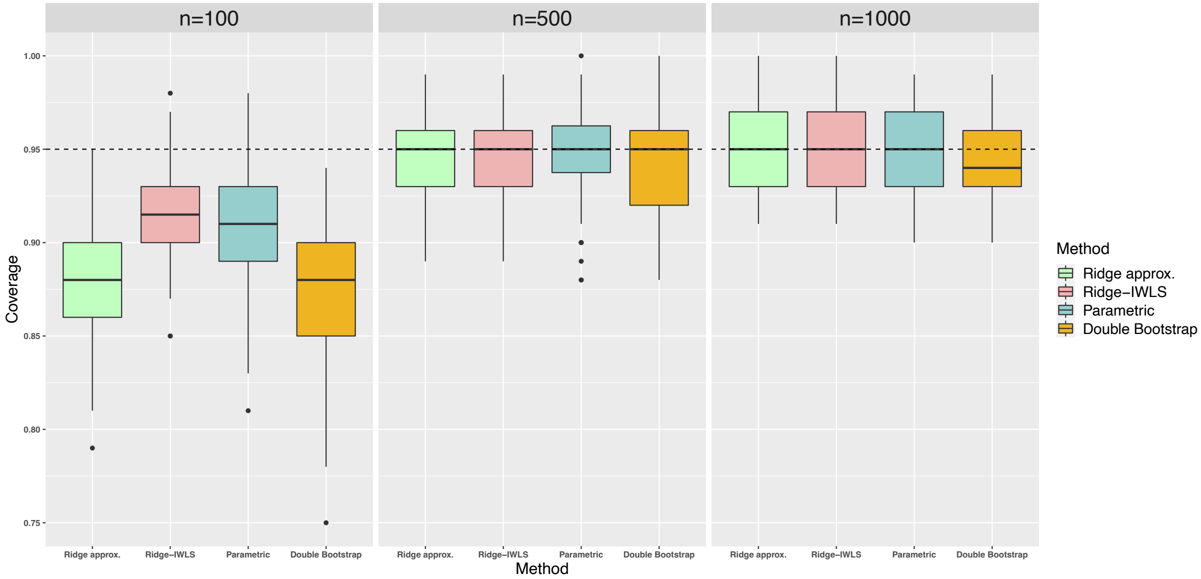

**Figure 4.1:** Coverage of 95% AFDB ridge approximation, AFDB ridge-IWLS, parametric and double bootstrap LASSO confidence regions for homoscedastic data

**Table 4.2:** Mean and standard deviation of the coverages of 95% AFDB ridge approximation, AFDB ridge-IWLS, parametric and double bootstrap LASSO confidence regions for homoscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| AFDB ridge approx | 0.881 (0.032) | 0.945 (0.024) | 0.949 (0.023) |
| AFDB ridge-IWLS | 0.917 (0.025) | 0.947 (0.022) | 0.949 (0.023) |
| Parametric | 0.9082 (0.031) | 0.947 (0.024) | 0.947 (0.021) |
| Double Bootstrap | 0.870 (0.037) | 0.941 (0.027) | 0.945 (0.021) |

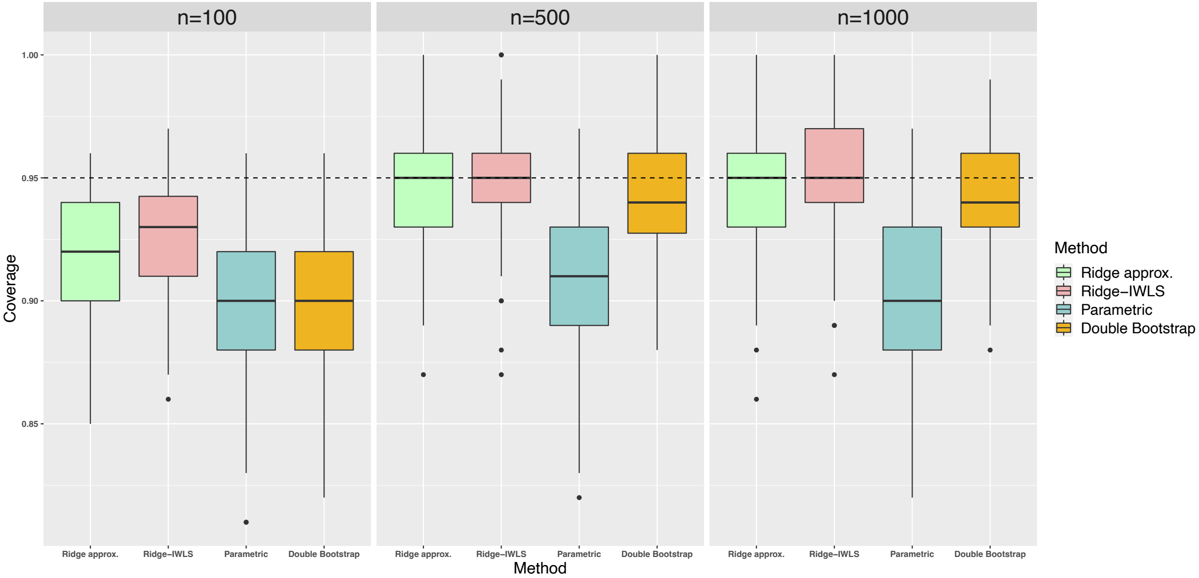

**Figure 4.2:** Coverage of 95% AFDB ridge approximation, AFDB ridge-IWLS, parametric and double bootstrap LASSO confidence regions for heteroscedastic data

**Table 4.3:** Mean and standard deviation of the coverages of 95% AFDB ridge approximation, AFDB ridge-IWLS, parametric and double bootstrap LASSO confidence regions for heteroscedastic data

| Method | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|
| AFDB ridge approx | 0.917 (0.027) | 0.946 (0.023) | 0.946 (0.026) |
| AFDB ridge-IWLS | 0.926 (0.024) | 0.948 (0.023) | 0.950 (0.024) |
| Parametric | 0.8976 (0.031) | 0.907 (0.031) | 0.9043 (0.031) |
| Double Bootstrap | 0.903 (0.03) | 0.94 (0.025) | 0.944 (0.023) |

## 4.6 Example

Recall Table 4.1 in which, for $\beta = (3, 0, -2, 0.5, 2.5)'$, we compared the estimation of coefficients across three approaches: LASSO, the ridge approximation and ridge-IWLS. In this case, using LASSO regression it was estimated that $\hat{\beta}_1^{lasso} = 2.6$. Suppose we wished to obtain a confidence interval for one of the regression coefficients, say $\beta_1$. To do this we can apply Theorem 4.4.2, taking $c$ to be the contrast vector $c = (1, 0, 0, 0, 0)$. In Figure 4.3 and Table 4.4, we compare 95% confidence intervals generated by the ridge approximation, ridge-IWLS and double bootstrap. All confidence intervals are fairly consistent with each other.

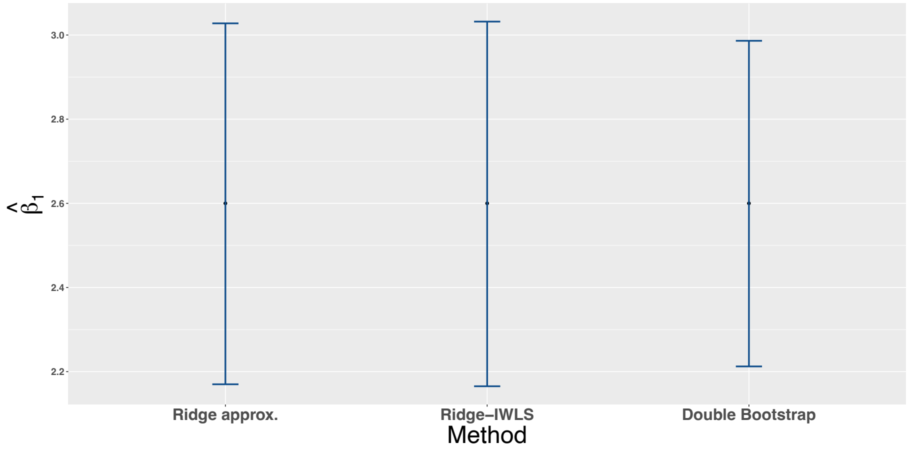

**Figure 4.3:** Comparing 95% confidence intervals in LASSO regression for the ridge approximation, ridge-IWLS and double bootstrap

**Table 4.4:** Comparing 95% confidence intervals in LASSO regression for the ridge approximation, ridge-IWLS and double bootstrap

| Method | Confidence Interval |
|---|---|
| Ridge Approximation | (2.169978, 3.027786) |
| Ridge-IWLS | (2.165282, 3.031926) |
| Double Bootstrap | (2.212433, 2.986302) |

## 4.7 Discussion

In this chapter, we showed how to approximate LASSO solutions using a ridge approximation. Alternatively we introduce the ridge-IWLS approximation as a preferred approach to the ridge approximation alone. Using either of these methods to estimate LASSO coefficients, the double bootstrap can then be applied as a means to estimate the sampling distribution and obtain confidence regions. To avoid the computational demand of the double bootstrap, the AFDB is illustrated in the context of LASSO regression. Through simulations, we can see that the AFDB performs analogously to the double bootstrap in a much more efficient manner. The choice of whether to adopt the AFDB or double bootstrap may depend on the size of the data and other computational constraints. Although we work in the setting where $n > p$ to avoid rank deficiency, penalized regression is still of interest in a variety of situations when $n > p$ like in the case of data that exhibits multicollinearity. For a discussion on confidence regions in high-dimensional models where $p > n$ see Van de Geer et al. (2014).

# Chapter 5

# Bootstrapping Generalized Linear Models

## 5.1 Introduction

This section references McCullagh and Nelder (1989). Generalized linear models (GLMs) are composed of three main components:

1. A response variable $Y$ whose components share a distribution in the exponential family.
2. A systematic component $\eta = X\beta$,

   where $X$ is the $n \times p$ matrix of explanatory variables and $\beta = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ vector of parameters.
3. A link function $g$ such that

   $$g(\mu) = \eta,$$

   where $\mu = \mathbb{E}[Y]$ and $g$ is a monotonic differentiable function.

A random variable $Y$ belongs to the exponential family if its density function is of the form

$$f_Y(y; \theta, \phi) = exp\left\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\right\}$$

for functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. These functions depend on the distribution of $Y$. $\theta$ is called the canonical parameter and $\phi$ is referred to as the dispersion parameter.

The log-likelihood function is given by

$$l(\theta, \phi; y) = (y\theta - b(\theta))/a(\phi) + c(y, \phi).$$

The mean and variance of $Y$ are then derived to be

$$\mathbb{E}(Y) = \mu = b'(\theta)$$

and

$$\text{Var}(Y) = b''(\theta)a(\phi) = V(\mu)a(\phi),$$

respectively. The normal, binomial and Poisson distribution are three commonly used distributions in the exponential family. Below is a table providing some characteristics for these distributions with the canonical link where $\theta = \eta = g(\mu) = X\beta$.

**Table 5.1:** Characteristics of the normal, binomial and Poisson Distributions

| | Normal | Binomial | Poisson |
|---|---|---|---|
| Notation | $\mathcal{N}(\mu, \sigma^2)$ | $B(m, \pi)/m$ | $P(\mu)$ |
| Link | identity | logit | log |
| $\mu(\eta)$ | $\eta$ | $e^\eta/(1 - e^\eta)$ | $e^\eta$ |
| $\eta = g(\mu)$ | $\mu$ | $\log\left(\frac{\mu}{1-\mu}\right)$ | $\log(\mu)$ |
| $V(\mu)$ | 1 | $\mu(1 - \mu)$ | $\mu$ |
| $a(\phi)$ | $\sigma^2$ | $1/m$ | 1 |

For GLMs, we can obtain maximum-likelihood estimates of $\beta$ by using an algorithm called iteratively weighted least squares (IWLS). The $(k+1)^{th}$ parameter estimate of $\beta$ can be solved as

$$\hat{\beta}^{(k+1)} = (X'W^{(k)}X)^{-1}X'W^{(k)}z^{(k)},$$

where $W^{(k)}$ is the $k^{th}$ iteration of the diagonal weight matrix $W$ with entries $w_{ii} = [\text{Var}(Y_i)]^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$ and $z^{(k)}$ is the $k^{th}$ iteration of the vector of adjusted responses $z$ composed of elements $z_i = \eta_i + (y_i - \hat{\mu}_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$. The IWLS algorithm is repeated until parameter convergence.

Typically, inference for the parameters in GLMs is based on quantities such as the Wald, Score or likelihood-ratio statistics (Dobson and Barnett, 2008).

The Wald statistic is given by

$$(\hat{\beta} - \beta)'\mathcal{I}(\beta)(\hat{\beta} - \beta),$$

where $\mathcal{I}(\beta)$ is the information matrix. Asymptotically, the Wald statistic follows a $\chi^2$ distribution with $p$ degrees of freedom. If we are interested in only one coefficient, we have

$$\hat{\beta} \sim \mathcal{N}(\beta, \mathcal{I}^{-1}).$$

The Wald statistic can then be used to perform hypothesis testing and obtain confidence intervals and regions for the parameters in a GLM. For instance, a confidence interval for $\beta$ could be computed as $\hat{\beta} \pm z_{\alpha/2} \times SE[\hat{\beta}]$. Of course, to guarantee exact results, we assume that the response variable is normally distributed which is not always the case (Dobson and Barnett, 2008). While alternative approaches based on quantities such as the Score or likelihood-ratio statistic require slightly weaker assumptions, they still rely on asymptotic convergence to achieve optimal results. This is why the consideration of nonparametric ways of conducting statistical inference for GLMs is important.

The rest of this chapter is organized as follows. Section 5.2 discusses both the one-step residual and wild bootstrap as nonparametric approaches to estimate a sampling distribution for the coefficients in GLMs. Simulation results are presented along with a real data example on data from a blood transfusion centre (Yeh et al., 2009; Dua and Graff, 2017). Section 5.3 develops the Analytic Wild Bootstrap (ANWB) methodology in the context of GLMs. Similarly, simulation results and a data application are also demonstrated. In Section 5.4, the case of overdispersed count data is introduced and the advantages of nonparametric approaches are highlighted in simulations and a data example using data pertaining to measurements on squids (Zuur et al., 2007). Finally, a discussion summarizing the key findings from the chapter is provided in Section 5.5.

## 5.2 Bootstrapping GLMs

### 5.2.1 Residual resampling

The bootstrap is a nonparametric method for conducting statistical inference in GLMs and it is an alternative to parametric options based on quantities such as the Wald and likelihood-ratio statistic. Moulton and Zeger (1991) developed a one-step bootstrapping procedure for the coefficients in GLMs based on either residual or vector resampling. As noted by Moulton and Zeger (1991), verifying distributional assumptions may not always be possible so the consideration of nonparametric approaches is important. The one-step procedure estimates bootstrap coefficients by using one iteration of IWLS. This approach is much more efficient than the alternative of estimating coefficients via IWLS a large number of times.

When implementing residual resampling for GLMs, one must decide on a choice of residuals. There are numerous types of residuals to consider such as the deviance, Pearson or standardized Pearson residuals. As described in McCullagh and Nelder (1989), the deviance, $D$, is a measure of how well a model fits the data, so each observation contributes a quantity $d_i$ such that $D = \sum_{i=1}^{n} d_i$. The deviance residuals are then computed as

$$r_i = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where $y_i$ is the observed value of the response variable and $\hat{\mu}_i$ is the predicted value of the response variable. Alternatively, the Pearson residuals are defined as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v_i}},$$

where $v_i$ is the estimated variance. Further standardizing by the diagonal entries of the hat matrix, the standardized Pearson residuals are calculated as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v_i(1-h_i)}},$$

where $h_i$ is the leverage of the $i^{th}$ observation computed from the $i^{th}$ diagonal entry of the hat matrix $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$. As suggested by Moulton and Zeger (1991), we use the standardized Pearson residuals as they are preferred from the standpoint of exchangeability. The computation of the hat matrix $H$ may not be desirable from a computation standpoint, especially as the dimensions of the data become large. Thus, it is possible to replace the scaling factors $1 - h_i$ with their average $1 - p/n$ (Friedl, 1997). These quantities have been shown to perform in an analogous manner through simulation studies (Friedl and Tilg, 1995). However, in this thesis we scale by the factor $1 - h_i$.

We describe the one-step residual resampling procedure for GLMs below as seen in Sartori (2011).

**Algorithm 5** One-step Residual Resampling Bootstrap

1. Using IWLS, estimate coefficients $\hat{\beta}$ and predicted responses $\hat{\mu}_i$.

2. Compute the standardized Pearson residuals $r_i$ and from them compute the mean-adjusted residuals $\bar{\varepsilon}_i = r_i - \bar{r}_i$.

3. Draw a sample with replacement of size $n$ from the mean-adjusted residuals and call it $\varepsilon^*$.

4. Estimate one-step bootstrap coefficients
   $\hat{\beta}^* = \hat{\beta} + (G'G)^{-1}G'\varepsilon^*$,
   where $G = W^{1/2}X$.

5. Repeat steps 3 and 4 $B$ times.

### 5.2.2 Wild bootstrap

While work has been done studying residual and vector resampling bootstrap for GLMs, the wild bootstrap has not been thoroughly explored. As an alternative to the one-step residual resampling bootstrap of Moulton and Zeger (1991), we will also investigate the one-step wild bootstrap's performance for a variety of GLMs.

**Algorithm 6** One-step Wild Bootstrap

```
1. Using IWLS, estimate coefficients $\hat{\beta}$ and predicted responses $\hat{\mu}_i$.

2. Compute the standardized Pearson residuals $r_i$.

3. Calculate $\varepsilon_i^* = \delta_i r_i$, where $\delta_i$ has $\mathbb{E}[\delta_i] = 0$ and $\text{Var}[\delta_i] = 1$.

4. Estimate one-step bootstrap coefficients
     $\hat{\beta}^* = \hat{\beta} + (G'G)^{-1}G'\varepsilon^*$,
   where $G = W^{1/2}X$.

5. Repeat steps 3 and 4 $B$ times.
```

In Moulton and Zeger (1991), the variance of the bootstrap coefficient is derived in the case of residual resampling. We derive the variance of the wild bootstrap coefficient for GLMs below.

**Proposition 1.** *The variance of the one-step wild bootstrap coefficient defined in Algorithm 6 is*

$$Var[\hat{\beta}^*] = (G'G)^{-1}G'RG(G'G)^{-1},$$

*where* $R = diag\left(\mathbb{E}[r_1^2], \ldots, \mathbb{E}[r_n^2]\right)$.

*Proof.* Consider $\hat{\beta}^*$, the coefficient vector from the one-step wild bootstrap defined in Algorithm 6. The variance can be written as

$$\text{Var}[\hat{\beta}^*] = (G'G)^{-1}G'\text{Var}[\varepsilon^*]G(G'G)^{-1}.$$

For diagonal entries where $i = j$, we have that

$$\begin{aligned}\text{Var}[\varepsilon_i^*] &= \text{Var}[\delta_i r_i] \\ &= \mathbb{E}[(\delta_i r_i - \mathbb{E}[\delta_i r_i])^2] \\ &= \mathbb{E}[\delta_i^2 r_i^2] \\ &= \mathbb{E}[r_i^2].\end{aligned}$$

Here, we used the fact that $\mathbb{E}[\delta_i] = 0$ and $\mathbb{E}[\delta_i^2] = 1$. Also, for $i \neq j$,

$$\begin{aligned}\text{Var}[\varepsilon^*]_{ij} &= \text{Cov}[\delta_i r_i, \delta_j r_j] \\ &= \mathbb{E}[\delta_i \delta_j r_i r_j] \\ &= 0.\end{aligned}$$

Thus,

$$\text{Var}[\hat{\beta}^*] = (G'G)^{-1}G'RG(G'G)^{-1},$$

where $R = \text{diag}\left(\mathbb{E}[r_1^2], \ldots, \mathbb{E}[r_n^2]\right)$. □

Below is a small simulation demonstrating the similar behaviours of the one-step residual resampling and wild bootstrap. Consider data with $n = 1000$, $p = 3$ and $\beta = (1.5, 2, -0.5)$. Suppose we wished to predict a binary response using logistic regression and we were interested in estimating the distribution of $\beta_1 = 1.5$. Using $B = 1000$ bootstrap replications, we compare the performance of the one-step residual and wild bootstraps in Figure 5.1.

From the histograms, one can see that the distributions of the estimated bootstrap coefficients using the one-step residual resampling and wild bootstrap are very similar with respective means and standard deviations of 1.507 (0.256) and 1.505 (0.248). The dashed blue line on the histograms represents the estimated coefficient of $\hat{\beta}_1 = 1.498$. To compute 95% confidence intervals for $\beta_1$, we could take the 0.025 and 0.975 quantiles of each distribution yielding confidence intervals of $(0.988, 1.963)$ for the residual resampling bootstrap and $(1.045, 1.996)$ for the wild bootstrap.

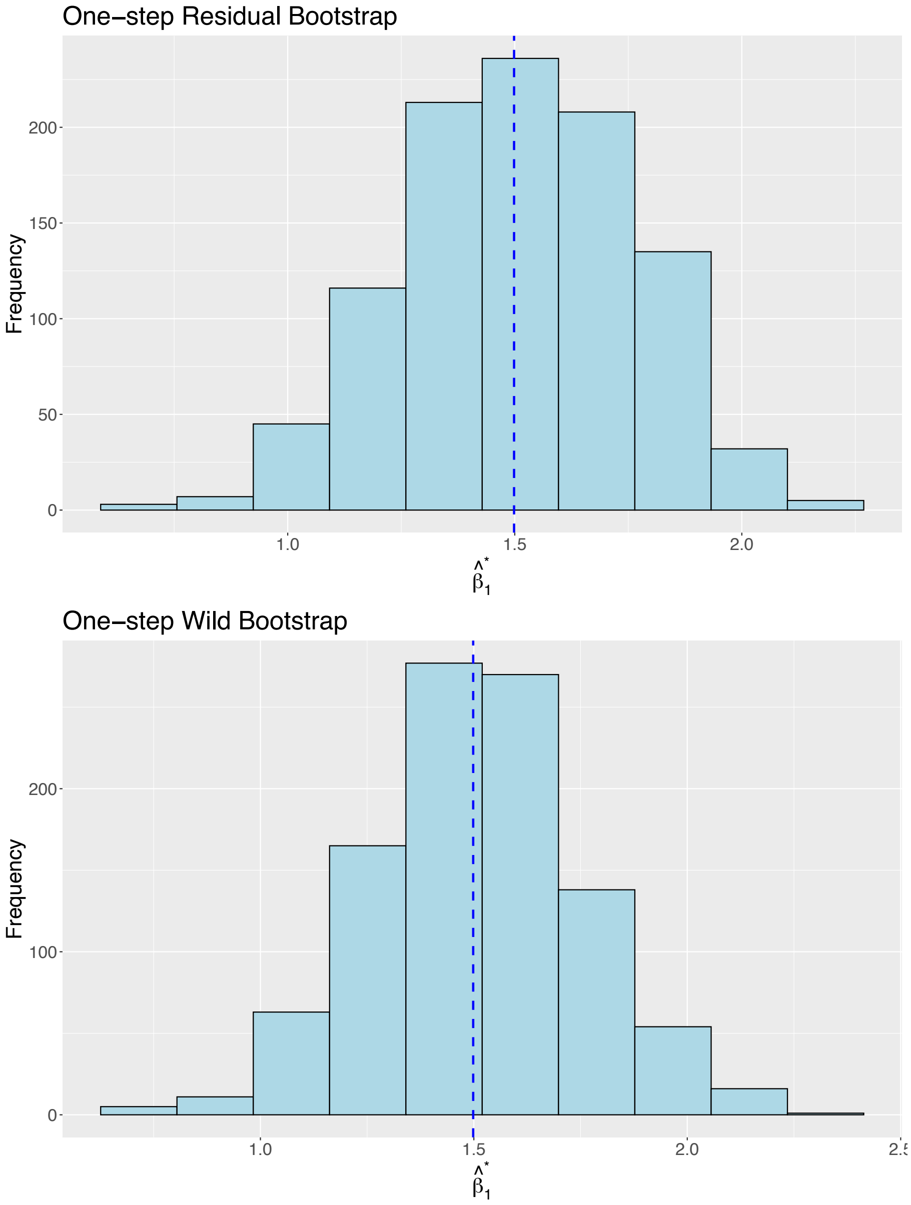

**Figure 5.1:** Comparison of the distributions of $\hat{\beta}_1$ generated by the one-step residual and wild bootstrap

### 5.2.3 Simulations

In this section, we look at simulations for both logistic and Poisson regression. Both scenarios consider $\alpha = 0.05$, $n = 1000$, $p = 3$ and $\beta$ selected from $U[-2, 2]$. For the bootstrap approaches, we utilize $B = 1000$ bootstrap replications. In this simulation, we are investigating the coverage of confidence interval's for $\beta$ using four difference approaches: two one-step bootstrap approaches, residual resampling and the wild bootstrap, and two parametric approaches based on the Wald statistic and likelihood-ratio statistic. We can see that all methods achieve the desired coverage level of 95%.

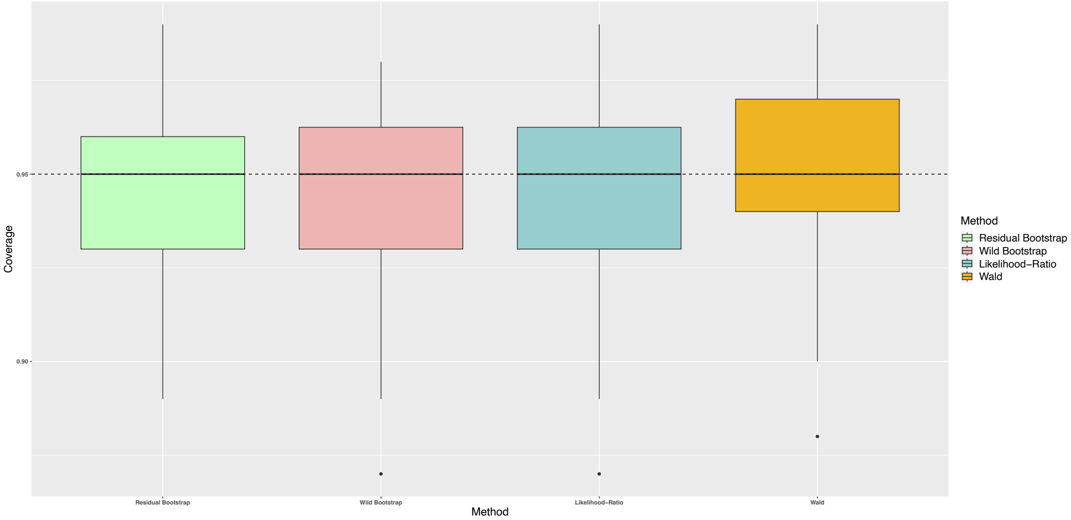

**Figure 5.2:** Coverage of various 95% confidence intervals in logistic regression

**Table 5.2:** Mean and standard deviation of the coverages of 95% confidence intervals for a logistic regression model

| Method | Coverage |
|---|---|
| Residual Bootstrap | 0.947 (0.021) |
| Wild Bootstrap | 0.946 (0.022) |
| Likelihood-Ratio | 0.947 (0.023) |
| Wald | 0.949 (0.021) |

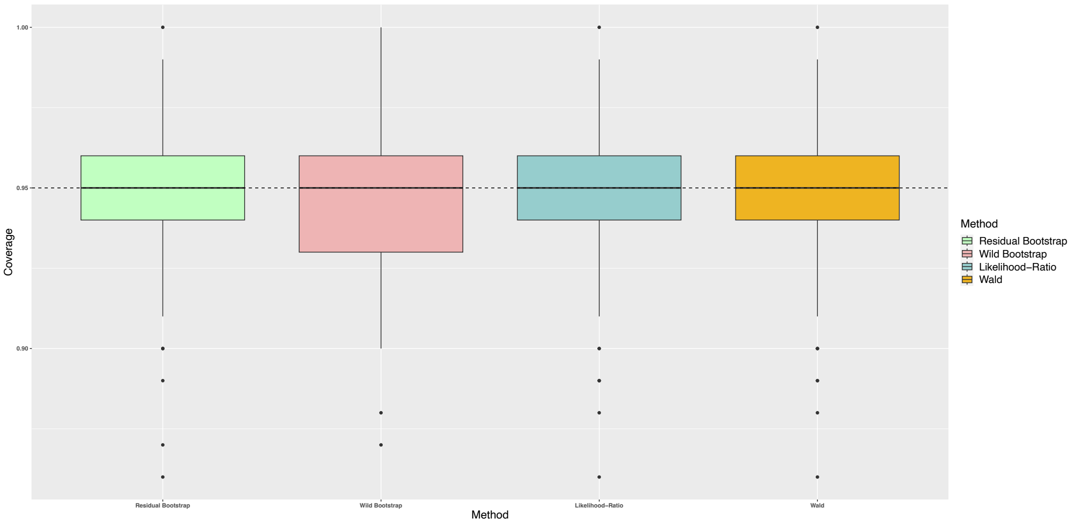

**Figure 5.3:** Coverage of various 95% confidence intervals in Poisson regression

**Table 5.3:** Mean and standard deviation of the coverages of 95% confidence intervals for a Poisson regression model

| Method | Coverage |
|---|---|
| Residual Bootstrap | 0.948 (0.025) |
| Wild Bootstrap | 0.945 (0.025) |
| Likelihood-Ratio | 0.947 (0.025) |
| Wald | 0.947 (0.024) |

`R` code has been provided in the Appendix to further illustrate the process of obtaining these confidence intervals in Poisson regression.

### 5.2.4 Data example

We will now illustrate the process of bootstrapping GLMs on a real data set. This data set was collected from a study on 748 randomly selected donors from a donor database of the Blood Transfusion Service Center in Hsin-Chu City in Taiwan (Yeh et al., 2009; Dua and Graff, 2017). The response variable of the study was a binary variable indicating whether or not the person donated blood (1-yes, 0-no). Covariates included recency in months since last donation ($X_1$), total number of donations ($X_2$), total blood donated in c.c. ($X_3$), and the time in months since first donation ($X_4$). Unsurprisingly, it was found that $X_2$ and $X_3$ were perfectly positively correlated with each other so only $X_2$ was considered in the model. After model selection, all variables were deemed important and the following logistic model was fit to the data.

$$\ln\left(\frac{p_{\text{donate}}}{1-p_{\text{donate}}}\right) = -0.45 - 0.10X_1 + 0.14X_2 - 0.02X_4,$$

where $p_{\text{donate}}$ indicates the probability of a person donating blood. For instance, for every additional donation made, a person is $e^{0.14} = 1.15$ times more likely to donate blood on average. Reported in Table 5.4 are the 95% confidence intervals for the coefficients in the above logistic regression model obtained by the following four methods: two one-step bootstrap approaches, residual resampling and the wild bootstrap, and two parametric approaches based on the Wald statistic and likelihood-ratio statistic.

**Table 5.4:** 95% confidence intervals for the blood transfusion data set

| Coefficient | Method | Confidence Interval |
|---|---|---|
| $\beta_0$ | Residual Bootstrap | $(-0.801, -0.097)$ |
| | Wild Bootstrap | $(-0.802, -0.109)$ |
| | Likelihood-Ratio | $(-0.806, -0.098)$ |
| | Wald | $(-0.803, -0.096)$ |
| $\beta_1$ | Residual Bootstrap | $(-0.133, -0.064)$ |
| | Wild Bootstrap | $(-0.133, -0.064)$ |
| | Likelihood-Ratio | $(-0.134, -0.066)$ |
| | Wald | $(-0.133, -0.065)$ |
| $\beta_2$ | Residual Bootstrap | $(0.084, 0.186)$ |
| | Wild Bootstrap | $(0.085, 0.191)$ |
| | Likelihood-Ratio | $(0.087, 0.188)$ |
| | Wald | $(0.085, 0.186)$ |
| $\beta_3$ | Residual Bootstrap | $(-0.035, -0.010)$ |
| | Wild Bootstrap | $(-0.035, -0.011)$ |
| | Likelihood-Ratio | $(-0.035, -0.012)$ |
| | Wald | $(-0.035, -0.011)$ |

## 5.3 Analytic Wild Bootstrap for GLMs

### 5.3.1 Confidence regions

For a generalized linear model, consider wild bootstrap coefficients

$$\hat{\beta}^* = (X'WX)^{-1}X'Wz^*,$$

where $z^* = X\hat{\beta} + \Gamma V^{1/2}\varepsilon^*$ and $\Gamma V^{1/2} = W^{-1/2}$.

Then we can write

$$\begin{aligned}\hat{\beta}^* &= (X'WX)^{-1}X'WX\hat{\beta} + (x'WX)^{-1}X'W\Gamma V^{1/2}\varepsilon^* \\ &= \hat{\beta} + (x'WX)^{-1}X'W\Gamma V^{1/2}\varepsilon^* \\ &= \hat{\beta} + (G'G)^{-1}G'\varepsilon^*,\end{aligned}$$

where $G = W^{1/2}X$. Thus, in an analogous manner to the case of least squares regression, we can write the bootstrap coefficient as a function of the estimated coefficient plus some linear operator applied to the perturbed residuals. Now we will introduce a theorem detailing an analytic way to consider confidence regions for generalized linear models.

**Theorem 5.3.1.** *For a generalized linear model $g(\mu) = X\beta$, denote $\hat{\beta}$ as the IWLS estimator of $\beta$ and consider the one-step wild bootstrap coefficients outlined in Algorithm 6: $\hat{\beta}^* = \hat{\beta} + (G'G)^{-1}G'\varepsilon^*$, where $G = W^{1/2}X$. Furthermore, assume that for $p \geq 2$ and for a constant $B_p$,*

$$\left(\mathbb{E}_\delta \left|\sum_{i=1}^n \delta_i r_i\right|^p\right)^{1/p} \leq B_p \left(\mathbb{E}_\delta \left|\sum_{i=1}^n \delta_i r_i\right|^2\right)^{1/2}$$

*for each sequence $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$ where $r_i$ are the standardized Pearson residuals and $\delta_i$ is a symmetric random variable with $\mathbb{E}[\delta_i] = 0$ and $Var[\delta_i] = 1$ .*

*Then, for* $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,

$$P\left((\hat{\beta}^* - \hat{\beta})' X' W X (\hat{\beta}^* - \hat{\beta}) \leq -\log(\alpha^*)\, 4C \left[\sum_{i,j=1}^{n} h_{ij,GLM}^2 r_i^2 r_j^2\right]^{1/2}\right) \geq 1 - \alpha,$$

*where* $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ *is the regularized incomplete beta function,* $\theta_1$ *and* $\theta_2$ *are fixed unknown constants,* $C$ *is a universal constant and* $h_{ij,GLM}$ *is the* $(i,j)^{th}$ *element of* $H_X = X(X'WX)^{-1}X'$.

*Proof.* Consider the following quadratic form

$$\begin{aligned}
(\hat{\beta}^* - \hat{\beta})'(X'WX)(\hat{\beta}^* - \hat{\beta}) &= \varepsilon^{*'} G(G'G)^{-1}(X'WX)(G'G)^{-1}G'\varepsilon^* \\
&= \varepsilon^{*'} W^{1/2} X(X'WX)^{-1}(X'WX)(X'WX)^{-1}X'W^{1/2}\varepsilon^* \\
&= \varepsilon^{*'} W^{1/2} X(X'WX)^{-1}X'W^{1/2}\varepsilon^* \\
&= \varepsilon^{*'} G(G'G)^{-1}G'\varepsilon^* \\
&= \varepsilon^{*'} H_{\text{GLM}}\varepsilon^* \\
&= \sum_{i,j} h_{ij,\text{GLM}}\varepsilon_i^*\varepsilon_j^* \\
&= \sum_{i,j} h_{ij,\text{GLM}}\delta_i\delta_j r_i r_j \\
&= T^2.
\end{aligned}$$

Thus, the rest of the proof follows from Theorem 2.5.1.

□

In comparison to other nonparametric approaches such as the bootstrap, this analytic method is superior from a computational standpoint. For example, using a quad-core processor obtaining confidence regions for the coefficients of a logistic glm for a data set with $n = 1000$, $p = 10$ and $B = 10000$ bootstrap replications takes approximately take 24.852 seconds, while using the analytic approach takes only 0.982 seconds.

## 5.3.2 Linear contrasts

In this section, we introduce an extension of Theorem 5.3.1 to derive confidence regions for the linear contrast $c'(\hat{\beta}^*(\lambda) - \hat{\beta}(\lambda))$.

**Theorem 5.3.2.** *Under the same settings as Theorem 5.3.1, let $c \in \mathbb{R}^p$ be a fixed contrast vector and $v_\lambda = (v_{1,\lambda}, \ldots, v_{n,\lambda})' = c'(X'WX)^{-1}X'$. Furthermore, assume that for a constant $B_p$,*

$$\left[\mathbb{E}_\delta \left|\sum_{i=1}^{n} v_{i,\lambda}\delta_i r_i\right|^p\right]^{1/p} \leq B_p^2 \left[\sum_{i=1}^{n} v_{i,\lambda}^2 r_i^2\right]^{1/2}$$

*for each sequence $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n \in \mathbb{R}$. Then, for $B_{2p} \propto [(2p)!/2^p p!]^{1/2p}$,*

$$P\left(\left|c'(\hat{\beta}^* - \hat{\beta})\right| \leq \left[-\log\left(\alpha^*\right) 4 \left[\sum_{i=1}^{n} v_{i,\lambda}^2 r_i^2\right]^{1/2}\right]^{1/2}\right) \geq 1 - \alpha,$$

*where $\alpha^* = I\left(\alpha\left[e/\sqrt{\pi} + e/\pi\right]^{-1}; \theta_1, \theta_2\right)$, $I(\alpha; \theta_1, \theta_2)$ is the regularized incomplete beta function and $\theta_1$, $\theta_2$ are fixed unknown constants.*

*Proof.* Consider the linear contrast

$$\begin{aligned} c'(\hat{\beta}^* - \hat{\beta}) &= c'(X'WX)^{-1}X'\varepsilon^* \\ &= v_\lambda'\varepsilon^* \\ &= \sum_{i=1}^{n} v_{i,\lambda}\varepsilon_i^* \\ &= \sum_{i=1}^{n} v_{i,\lambda}\delta_i r_i \\ &= T, \end{aligned}$$

where $v_\lambda = (v_{1,\lambda}, \ldots, v_{n,\lambda})' = c'(X'WX)^{-1}X'$. The proof follows the same procedure as seen in the proof of Theorem 2.5.2. □

### 5.3.3 Simulations

This section demonstrates simulations for logistic and Poisson regression. Both scenarios consider $\alpha = 0.05$, $n = 1000$, $p = 3$ and $\beta$ selected from $U[-3, 3]$. For the bootstrap approaches we utilized $B = 1000$ bootstrap replications. We are exploring the performance of confidence regions of two one-step bootstrap approaches: the wild bootstrap and the analytic wild bootstrap as well as confidence regions based on the Wald statistic. While the methods do well in terms of coverage, the analytic approach takes dramatically less time to compute.

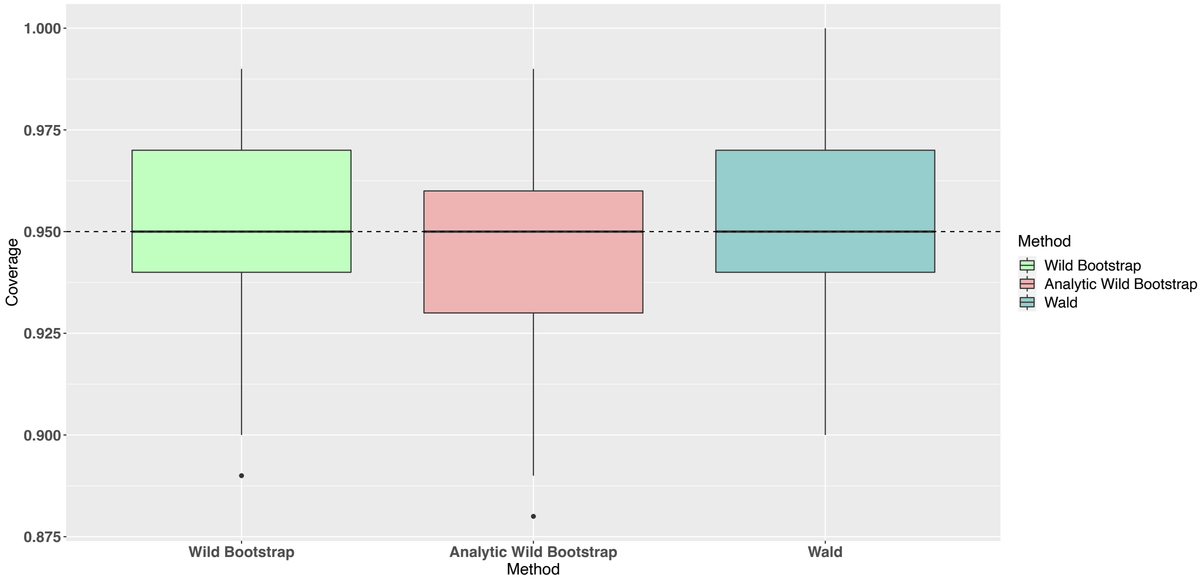

**Figure 5.4:** Coverage of various 95% confidence intervals in logistic regression

**Table 5.5:** Mean and standard deviation of the coverages of 95% confidence intervals for a logistic regression model

| Method | Coverage |
|---|---|
| Wild Bootstrap | 0.948 (0.021) |
| Analytic Wild Bootstrap | 0.950 (0.022) |
| Wald | 0.951 (0.022) |

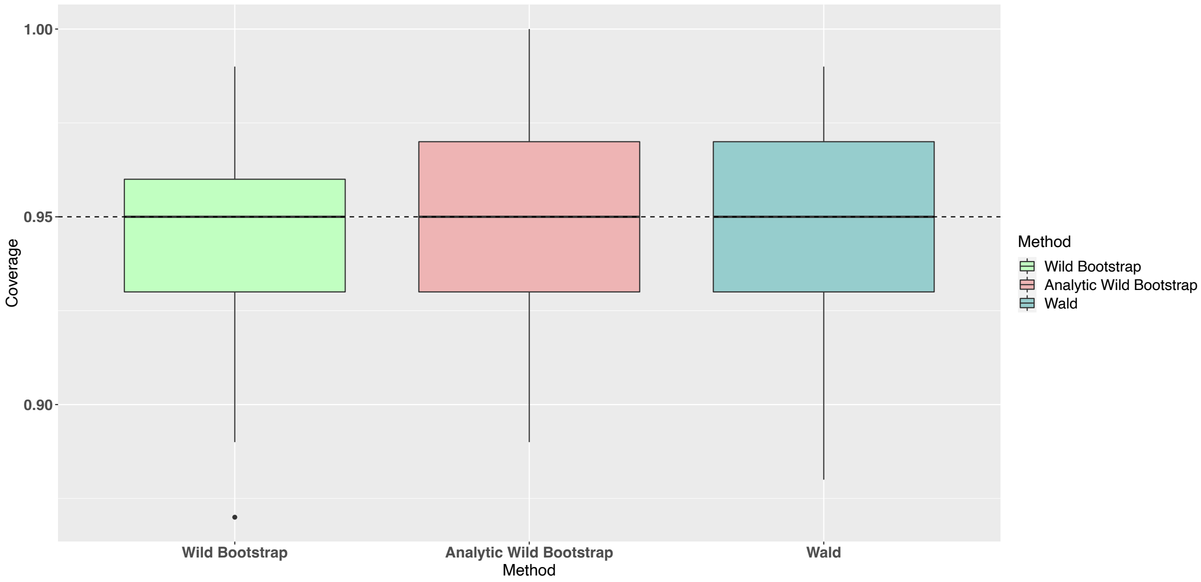

**Figure 5.5:** Coverage of various 95% confidence regions in Poisson regression

**Table 5.6:** Mean and standard deviation of the coverages of 95% confidence regions for a Poisson regression model

| Method | Coverage |
|---|---|
| Wild Bootstrap | 0.947 (0.024) |
| Analytic Wild Bootstrap | 0.950 (0.025) |
| Wald | 0.949 (0.024) |

### 5.3.4 Data example

Revisiting the blood transfusion data set from Section 5.2.4, we will demonstrate the application of the analytic wild bootstrap confidence regions using logistic regression. Recall that the response variable of the study was a binary variable indicating whether or not the person donated blood (1-yes, 0-no). Independent variables recorded included recency in months since last donation ($X_1$), total number of donations ($X_2$), total blood donated in c.c. ($X_3$), and the time in months since first donation ($X_4$). After model selection, the following logistic model was fit

$$\begin{aligned}\ln\left(\frac{p_{\text{donate}}}{1-p_{\text{donate}}}\right) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 \\ &= -0.45 - 0.10X_1 + 0.14X_2 - 0.02X_4.\end{aligned}$$

A plot of confidence regions for $\beta_1$ and $\beta_2$ computed using the analytic wild bootstrap, Wald and wild bootstrap approach are displayed below. All three methods produced very similar ellipses, with radii of 3.09, 3.05, and 3.08 for the ANWB, Wald and Wild bootstrap, respectively.

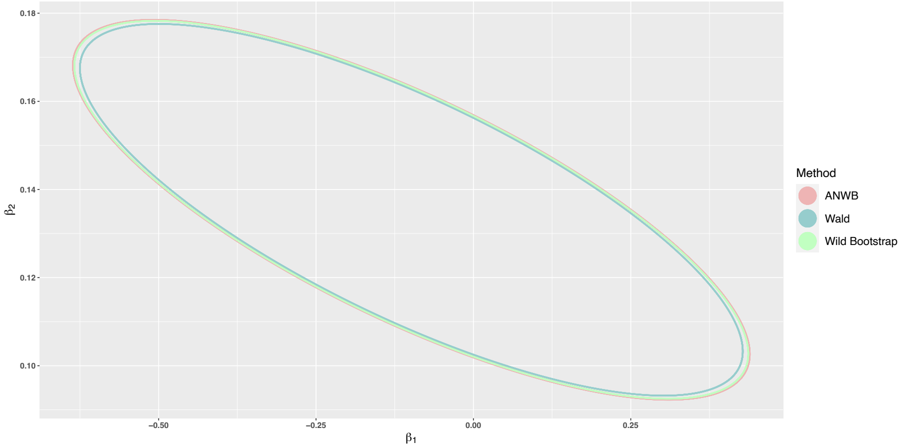

**Figure 5.6:** Coverage of various 95% confidence regions for the blood transfusion data set

## 5.4 Overdispersed Count Data

### 5.4.1 Overdispersion

In Poisson regression, a key assumption is that the mean and variance are equal. That is, $\mathbb{E}(Y) = \text{Var}(Y) = \mu$. However, there are cases when the variance is greater than the mean leading to a phenomenon called overdispersion. For overdispersed count data, the variance is proportional to the mean such that $\text{Var}(Y) = \phi\text{E}(Y)$, where $\phi > 1$ is the dispersion parameter. When $\phi = 1$, the problem reduces to the Poisson distribution. The issue of underdispersion where $\phi < 1$ can also be considered, but in this thesis we will primarily focus on the case where $\phi > 1$. As described by Zuur et al. (2009), overdispersion can arise in a variety of settings, such as when the observations are correlated or in the case of zero-inflated responses. On the other hand, they also note that there can be apparent overdispersion due to model misspecifications such as the exclusion of certain explanatory variables or interaction terms in the model, non-linearity and outliers. Figure 5.7 provides a visualization of overdispersed count data.

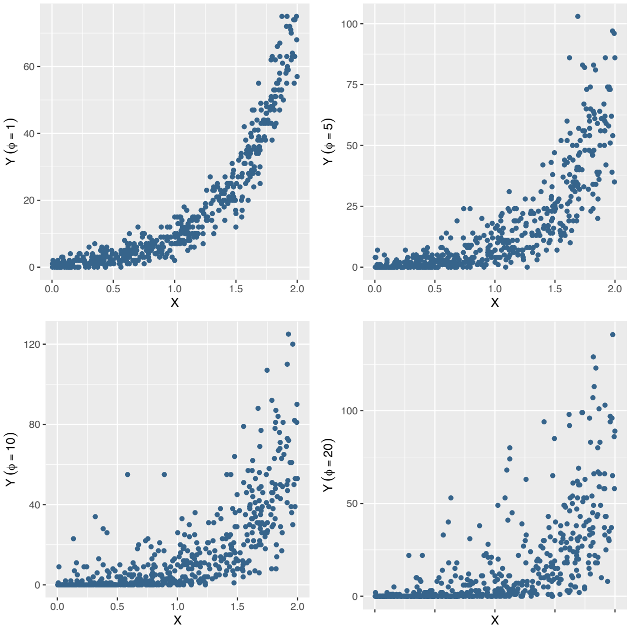

**Figure 5.7:** Overdispersed count data for $\phi \in \{1, 5, 10, 20\}$

To estimate the overdispersion parameter $\phi$, we can take the ratio of the residual deviance to its degree of freedom (Gail et al., 2007). That is,

$$\hat{\phi} = \frac{D}{n-p},$$

where $D$ is the residual deviance. Thus, $\hat{\phi} > 1$ could indicate overdispersed data which requires additional consideration when choosing a model to work with.

If overdispersion is present in count data, using the Poisson model without accounting for overdispersion can be misleading in terms of the resulting standard errors. The Poisson model may underestimate the standard errors, resulting in an erroneous significance of regression coefficients (Ismail and Jemain, 2007). A natural question is then how should we account for overdispersion in the model? The negative binomial and quasi-Poisson models are two common choices that can accommodate overdispersed count data. The key difference between these two models is the relationship between the variance and the mean. Using the notation from Zuur et al. (2009), for $Y \sim \mathrm{NB}(\mu, k)$, the negative binomial model specifies that

$$\mathbb{E}(Y) = \mu$$

and

$$\mathrm{Var}(Y) = \mu + \mu^2/k,$$

where $k^{-1}$ is the dispersion parameter. Thus, the variance is a quadratic function of the mean.

For the quasi-Poisson model, the variance is linearly related to the mean such that

$$\mathbb{E}(Y) = \mu$$

and

$$\mathrm{Var}(Y) = \phi\mu.$$

Ver Hoef and Boveng (2007) highlight that since both of these models assume a different variance, coefficient estimates will vary as the weights used in IWLS are inversely proportional to the variance. They conveniently provide the following comparison of the weight matrices for the two models. For the quasi-Poisson model we have that

$$W = \operatorname{diag}\left(\frac{\mu_1}{\phi}, \dots, \frac{\mu_n}{\phi}\right),$$

while for the negative binomial model the weight matrix is given by

$$W = \operatorname{diag}\left(\frac{\mu_1}{1 + k^{-1}\mu_1}, \dots, \frac{\mu_n}{1 + k^{-1}\mu_n}\right).$$

The advantages of the quasi-Poisson model are that it is a simple adjustment to the mean-variance relationship of the Poisson distribution and parameters are still easy to interpret (Ver Hoef and Boveng, 2007). Going forward, we will be discussing inferential techniques with regards to the quasi-Poisson model.

### 5.4.2 Simulations

Using the quasi-Poisson model, we may be interested in obtaining estimated confidence regions for regression parameters. When we are presented with overdispersed count data, we can turn to the wild bootstrap or ANWB as a means to construct confidence regions. In the simulations below, we consider the coverage of three approaches: the wild bootstrap, ANWB and Wald. We compare the methods across varying levels of overdispersion where $\phi = 1, 2, 3, 4$ and for sample sizes of $n = 100$ and $n = 1000$. Setting $\alpha = 0.05$, the desired level of coverage is 0.95. We can see that both the wild bootstrap and ANWB have median coverages close to the target of 0.95, while as $\phi$ increases it is apparent that the Wald approach performs progressively worse in terms of coverage.

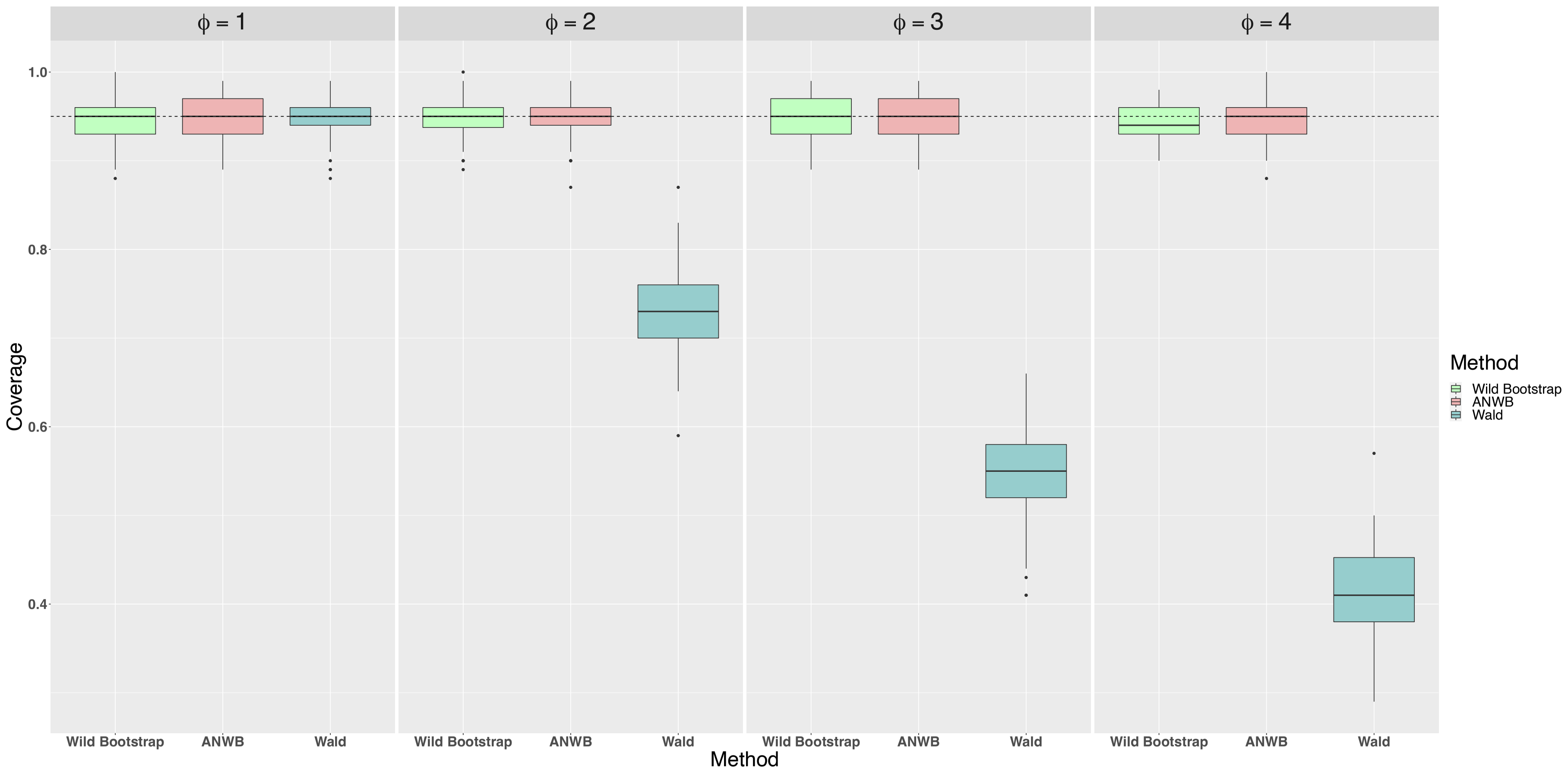

**Figure 5.8:** Coverage of various 95% confidence regions for overdispersed Poisson data with $n = 1000$

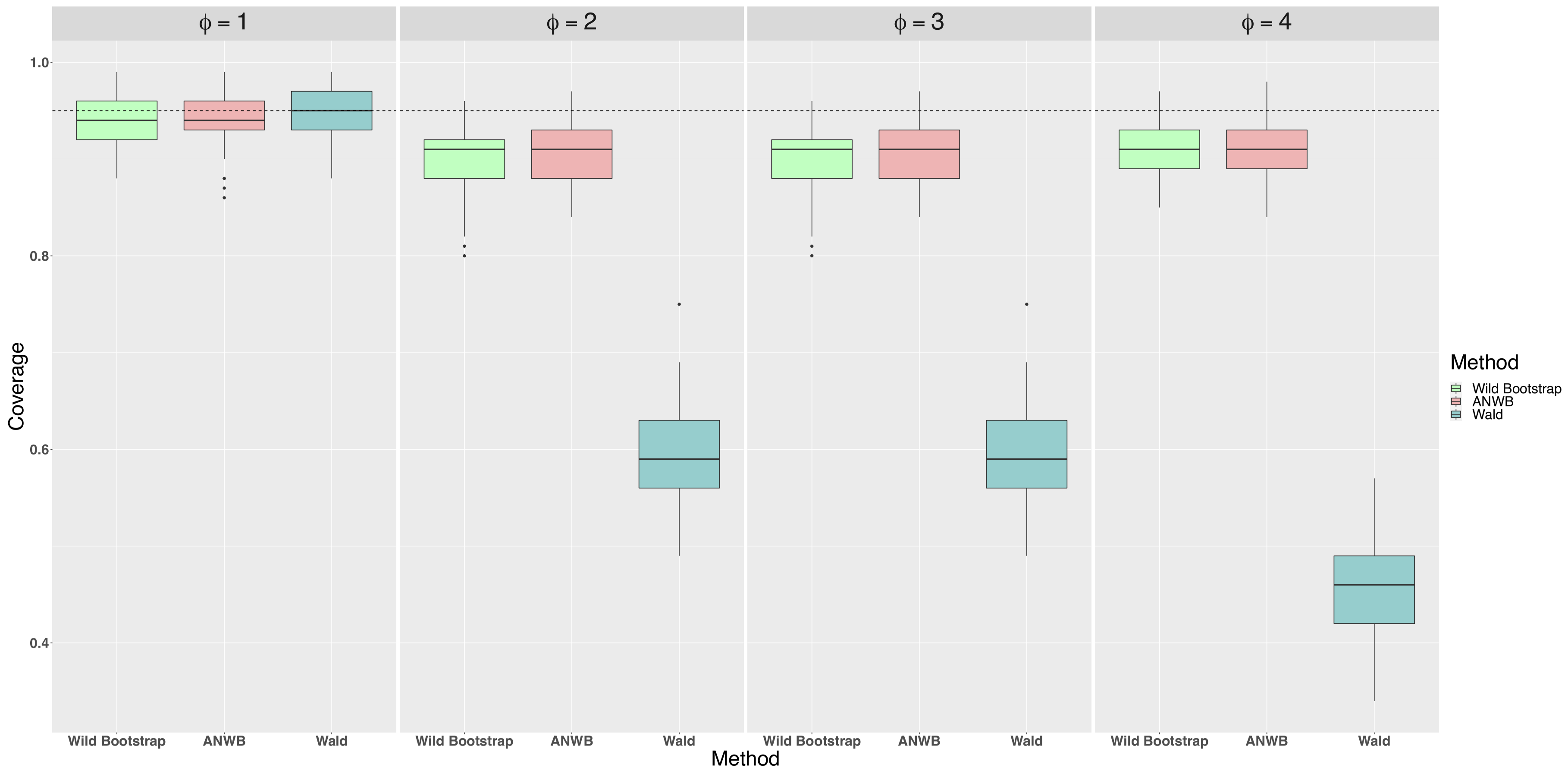

**Figure 5.9:** Coverage of various 95% confidence regions for overdispersed Poisson data with $n = 100$

**Table 5.7:** Mean and standard deviation of the coverages of 95% confidence regions for an overdispersed Poisson regression model with $n = 1000$

| Dispersion Parameter | Method | Coverage |
|---|---|---|
| $\phi = 1$ | Wild Bootstrap | 0.949 (0.023) |
| | ANWB | 0.948 (0.024) |
| | Wald | 0.950 (0.023) |
| $\phi = 2$ | Wild Bootstrap | 0.948 (0.022) |
| | ANWB | 0.949 (0.022) |
| | Wald | 0.727 (0.049) |
| $\phi = 3$ | Wild Bootstrap | 0.949 (0.023) |
| | ANWB | 0.949 (0.023) |
| | Wald | 0.546 (0.049) |
| $\phi = 4$ | Wild Bootstrap | 0.946 (0.024) |
| | ANWB | 0.947 (0.021) |
| | Wald | 0.415 (0.050) |

**Table 5.8:** Mean and standard deviation of the coverages of 95% confidence regions for an overdispersed Poisson regression model with $n = 100$

| Dispersion Parameter | Method | Coverage |
|---|---|---|
| $\phi = 1$ | Wild Bootstrap | 0.941 (0.024) |
| | ANWB | 0.943 (0.024) |
| | Wald | 0.949 (0.024) |
| $\phi = 2$ | Wild Bootstrap | 0.943 (0.024) |
| | ANWB | 0.945 (0.024) |
| | Wald | 0.733 (0.042) |
| $\phi = 3$ | Wild Bootstrap | 0.904 (0.031) |
| | ANWB | 0.900 (0.031) |
| | Wald | 0.594 (0.048) |
| $\phi = 4$ | Wild Bootstrap | 0.912 (0.030) |
| | ANWB | 0.910 (0.028) |
| | Wald | 0.458 (0.049) |

### 5.4.3 Data example

This section demonstrates the various approaches discussed to compute confidence regions for overdispersed count data on a real data set. The data set in question can be found in Zuur et al. (2007) and is concerned with measurements recorded on Dutch coastal areas for the purposes of assessing the impact climate change has had on local species (Janssen et al., 2007). The response variable is species richness which is a measure of biodiversity. Explanatory variables of interest include: NAP, the height of the sampling station relative to the mean tidal level; exposure, which is a nominal variable with three levels indicating the varying degrees of exposure to different coastal elements; and week, which has four levels corresponding to four weeks in June that samples were collected. Further details regarding the data set are discussed in Zuur et al. (2007) and Janssen et al. (2007).

First, we provide an overview of the discussion in Zuur et al. (2007) leading to the conclusion that the RIKZ data may be overdispersed. We initially consider a simple regression model with NAP as the only explanatory variable.

```
> summary(RIKZ_fit1)

Call:
glm(formula = Richness ~ NAP, family = poisson(link = "log"),
    data = RIKZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2029  -1.2432  -0.9199   0.3943   4.3256

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18

Number of Fisher Scoring iterations: 5
```

From this output, the ratio of the residual deviance (113.18) and its respective degrees of freedom (43) is clearly greater than 1 which may indicate overdispersion in the data. The overdispersion can further be seen by looking at a scatterplot of species richness and NAP in Figure 5.10.

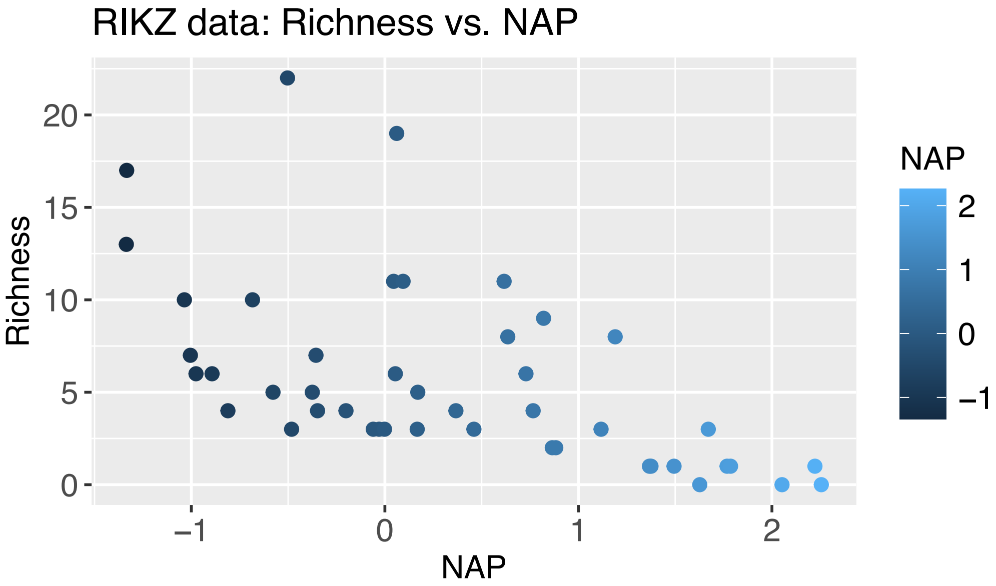

**Figure 5.10:** Scatterplot of Richness vs. NAP for RIKZ data

To adjust for the overdispersion in the model, we can use the quasi-Poisson model.

```
> summary(RIKZ_fit2)

Call:
glm(formula = Richness ~ NAP, family = quasipoisson(link = "log"),
    data = RIKZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2029  -1.2432  -0.9199   0.3943   4.3256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7910     0.1104  16.218  < 2e-16 ***
NAP          -0.5560     0.1250  -4.448 6.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 3.044178)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

From the R output above, we can see that although coefficient estimates remain the same, the standard errors have been increased. For instance, although the coefficient estimates for both models are $-0.556$, the standard error for the NAP coefficient was previously 0.07163 and has been multiplied by the square root of the estimated dispersion parameter $\sqrt{3.044178}$ to result in a new standard error of 0.125. Note that in R, the overdispersion parameter is estimated by dividing Pearson's chi-square statistic, $\chi^2$, by $n - p$.

As described in Zuur et al. (2007), the preferred model when considering overdispersion includes NAP and week as covariates.

```
> summary(RIKZ_fit3)

Call:
glm(formula = Richness ~ NAP + factor(week),
family = quasipoisson(link = "log"), data = RIKZ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9727  -0.5757  -0.1865   0.3577   2.7536

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.32603    0.11169  20.826  < 2e-16 ***
NAP           -0.44821    0.08288  -5.408 3.20e-06 ***
factor(week)2 -1.21144    0.21331  -5.679 1.33e-06 ***
factor(week)3 -0.80473    0.18091  -4.448 6.74e-05 ***
factor(week)4 -0.11102    0.22404  -0.496    0.623
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.284359)

    Null deviance: 179.753  on 44  degrees of freedom
Residual deviance:  53.466  on 40  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure 5.11 provides estimated 95% confidence regions obtained by the ANWB, Wald statistic and wild bootstrap. As expected from the simulations, confidence regions based on the Wald statistic are smaller than those obtained from the nonparametric approaches due to overdispersion in the data.

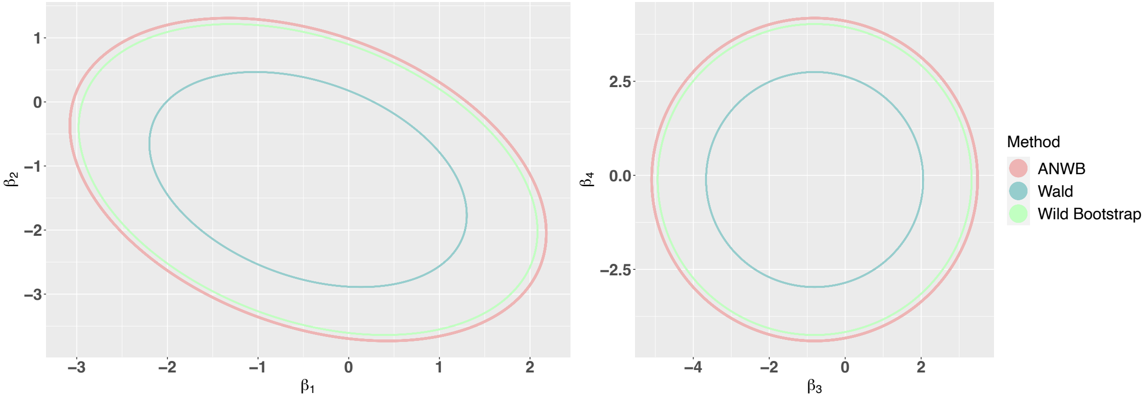

**Figure 5.11:** 95% confidence regions for RIKZ data set

## 5.5 Discussion

In this chapter, we discussed a variety of methods used to conduct statistical inference when dealing with generalized linear models. While approaches based on quantities such as the Wald, likelihood-ratio and Score statistic are commonly used, they are based on asymptotic results that may be difficult to verify. Looking at the simulation results in Section 5.2.3, the coverages of the confidence intervals computed in logistic and Poisson regression all achieve comparable results. Similarly, in 5.3.3 the confidence regions based upon the ANWB, wild bootstrap and Wald statistic all demonstrate similar coverage probabilities.

However, it is in Section 5.4 where the advantages of the bootstrap approaches over typical parametric methods are evident. When presented with overdispersion in Poisson regression, bootstrap methods are considerably better in terms of coverage. Moreover, in Figures 5.4 and 5.5, it is shown that as the dispersion parameter, $\phi$, increases, the performance of the Wald confidence regions dramatically declines while the other methods remain relatively unaffected. While the ANWB and wild bootstrap approaches exhibit similar performance in terms of coverage, one should keep in mind that the ANWB is a more computationally efficient process.

# Chapter 6

# Conclusion

In this chapter, we provide a summary of the thesis and suggest some potential future directions of research based upon this work.

## 6.1 Summary

In this thesis, a nonparametric framework for conducting statistical inference for a variety of linear models is introduced. Unlike many parametric techniques that rely on distributional assumptions and asymptotic results, our analytic method applies the concentration of measure phenomenon to the wild bootstrap for regression coefficients. In doing so, we retain the appealing nonparametric aspects of the bootstrap while overcoming its main criticism: computational inefficiency.

In Chapter 1, we provided an introduction and overview of the thesis and briefly explained the research problem. Chapter 2 first introduced the concept of concentration of measure and developed the analytic wild bootstrap in the context of least squares regression. In Chapter 3, we developed the methodology for the estimators in ridge regression, where typical bootstrapping is not necessarily appropriate. Chapter 4 provided extensions to LASSO regression where the estimated coefficients have no closed-form solution. In Chapter 5, we moved into the territory of generalized linear models and discussed a variety of cases, highlighting the applications on overdispersed Poisson data.

Ultimately, this work has provided an analytic mechanism for the computation of confidence regions and intervals for various linear models. Hopefully this thesis motivates the use of concentration on other computationally-intensive statistical processes. The following section proposes some possible prospective ideas to explore in this field of research.

## 6.2 Future Work

### 6.2.1 Penalized regression

In least squares regression, we must work in the setting where $n > p$ to avoid rank deficiency. In Chapter 4, we discuss an analytic solution for estimating confidence regions in the context of LASSO regression, where it is often of interest to look at the case where $p >> n$ for feature selection. Implementing penalized regression models is still pertinent in a variety of situations when $n > p$ (e.g., multicollinear data), but looking into how this methodology can accommodate high-dimensional data, in terms of the number of variables, would be a logical next step.

### 6.2.2 Dependent data

In this thesis, we assume that the errors are independently distributed. When working with time series data, this is not the case due to the temporal dependency. Block-based techniques for bootstrapping time series data have been studied such as the Moving Block Bootstrap (MBB) (Kunsch, 1989). With the MBB, the researcher defines a number of overlapping blocks and bootstrap samples are obtained for each block, preserving the autocorrelation and time series structure within the blocks (Radovanov and Marcikić, 2014). Other methods such as the dependent wild bootstrap (Shao, 2010) have also been investigated. As is the case with resampling schemes of this nature, they are typically slow from a computational standpoint. Thus, trying to approach bootstrapping time series data from an analytic perspective using concentration techniques could be of interest for future work in this area.

# Bibliography

Alves, C. and Sapozhnikov, A. (2019). Decoupling inequalities and supercritical percolation for the vacant set of random walk loop soup. *Electronic Journal of Probability*, 24:1–34.

Beran, R. (1986). Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1295–1298.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Buchholz, A. (2001). Operator khintchine inequality in non-commutative probability. *Mathematische Annalen*, 319(1):1–16.

Burak, K. L. and Kashlak, A. B. (2022). Nonparametric confidence regions via the analytic wild bootstrap. *Canadian Journal of Statistics*.

Chatterjee, A. and Lahiri, S. (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509.

Davidson, R. and MacKinnon, J. G. (2007). Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics & Data Analysis*, 51(7):3259–3281.

De, A., Diakonikolas, I., and Servedio, R. A. (2016). A robust khintchine inequality, and algorithms for computing optimal constants in fourier analysis and high-dimensional geometry. *SIAM Journal on Discrete Mathematics*, 30(2):1058–1094.

De la Pena, V. and Giné, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media.

Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dubhashi, D. P. and Panconesi, A. (2009). *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Efron, B. and Hastie, T. (2016). Computer age statistical inference: Algorithms, evidence, and data science.

Flachaire, E. (1999). A better way to bootstrap pairs. *Economics Letters*, 64(3):257–262.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis*, 49(2):361–376.

Floret, K. and Matos, M. C. (1995). Application of a khintchine inequality to holomorphic mappings. *Mathematische Nachrichten*, 176(1):65–72.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228.

Friedl, H. (1997). On the asymptotic moments of pearson type statistics based on resampling procedures. *Computational Statistics*, 12(2):265–278.

Friedl, H. and Tilg, N. (1995). Variance estimates in logistic regression using the bootstrap. *Communications in statistics-theory and methods*, 24(2):473–486.

Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., and Wong, W. (2007). Statistics for biology and health.

Garling, D. J. H. (2007). *Inequalities: A Journey into Linear Analysis*. Cambridge University Press.

Hagemann, A. (2017). Cluster-robust bootstrap inference in quantile regression models. *Journal of the American Statistical Association*, 112(517):446–456.

Havrilla, A. and Tkocz, T. (2020). Sharp khinchin-type inequalities for symmetric discrete uniform random variables.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J. (2017). Gender discrimination, version 1.

Ismail, N. and Jemain, A. A. (2007). Handling overdispersion with negative binomial and generalized poisson regression models. In *Casualty actuarial society forum*, volume 2007, pages 103–58. Citeseer.

Janssen, G., Mulder, S., Zuur, A., Ieno, E., and Smith, G. (2007). Univariate and multivariate analysis applied on a dutch sandy beach community. In *Analysing Ecological Data*, pages 485–501. Springer.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622.

Kashlak, A. B., Myroshnychenko, S., and Spektor, S. (2022). Analytic permutation testing for functional data anova. *Journal of Computational and Graphical Statistics*, pages 1–10.

Khintchine, A. (1923). Über dyadische brüche. *Mathematische Zeitschrift*, 18(1):109–116.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241.

Kwapien, S. (1987). Decoupling Inequalities for Polynomial Chaos. *The Annals of Probability*, 15(3):1062 – 1071.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon.* American Mathematical Society.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer Berlin Heidelberg, 1st edition.

Li, S. (2020). Debiasing the debiased lasso with bootstrap. *Electronic Journal of Statistics*, 14(1):2298–2337.

Liu, R. Y. (1988). Bootstrap Procedures under some Non-I.I.D. Models. *The Annals of Statistics*, 16(4):1696 – 1708.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics. *Economic Record*, 82:S2–S18.

Makarychev, K. and Sviridenko, M. (2018). Solving optimization problems with diseconomies of scale via decoupling. *Journal of the ACM (JACM)*, 65(6):1–27.

Mammen, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics*, 21(1):255 – 285.

McConnell, T. R. and Taqqu, M. S. (1986). Decoupling inequalities for multilinear forms in independent symmetric random variables. *The Annals of Probability*, 14(3):943–954.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition.* Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

McCullough, B. and Vinod, H. (1998). Implementing the double bootstrap. *Computational Economics*, 12(1):79–95.

Moulton, L. H. and Zeger, S. L. (1991). Bootstrapping generalized linear models.

Nguyen, N., Drineas, P., and Tran, T. (2009). Matrix sparsification via the khintchine inequality.

Nickerson, D. M. (1994). Construction of a conservative confidence region from projections of an exact confidence region in multiple linear regression. *The American Statistician*, 48(2):120–124.

Radovanov, B. and Marcikić, A. (2014). A comparison of four different block bootstrap methods. *Croatian Operational Research Review*, pages 189–202.

Sartori, S. (2011). Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235.

Spektor, S. (2016). Restricted khinchine inequality. *Canadian Mathematical Bulletin*, 59(1):204–210.

Talagrand, M. (1996). A new look at independence. *The Annals of probability*, pages 1–34.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567.

Van De Geer, S. (2019). On the asymptotic variance of the debiased lasso. *Electronic Journal of Statistics*, 13(2):2970–3008.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Vinod, H. (1987). Confidence intervals for ridge regression parameters. *Time series and econometric modelling*, pages 279–300.

Vinod, H. (1995). Double bootstrap for shrinkage estimators. *Journal of Econometrics*, 68(2):287–302.

Wang, L., Van Keilegom, I., and Maidman, A. (2018). Wild residual bootstrap inference for penalized quantile regression with heteroscedastic errors. *Biometrika*, 105(4):859–872.

Watson, G. N. (1959). A note on gamma functions. *Edinburgh Mathematical Notes*, 42:7–9.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295.

Yeh, I.-C., Yang, K.-J., and Ting, T.-M. (2009). Knowledge discovery on rfm model using bernoulli sequence. *Expert Systems with Applications*, 36(3):5866–5871.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zuur, A. F., Ieno, E. N., and Smith, G. M. (2007). Analysing ecological data.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009). *GLM and GAM for Count Data*. Springer New York, New York, NY.

# Appendix

R code for the computation confidence intervals in Poisson regression from Chapter 5. Note that parametric confidence intervals based on the Wald and likelihood-ratio statistic were computed using the `confint2` function in the `glmtoolbox` package.

```
# initialization
set.seed(40899)
n <- 1000
p <- 3
alpha <- 0.05
betas <- runif(p,-2,2)

X <- matrix(runif(n*p), nrow=n, ncol=p)
Xb <- X%*%betas
lambda <- exp(Xb)
y <- rpois(n, lambda)

# fit model
fit <- glm(y ~ X-1, family = poisson(link = "log"))
beta.hat <- coef(fit)

# compute standardized pearson residuals
mu.hat <- predict(fit, newx=X, type="response") # fitted values
v.hat <- fit$family$variance(mu.hat) # estimate variance
pearson.res <- (y - mu.hat)/sqrt(v.hat) # pearson residuals
hi <- hatvalues(fit) # hat values
```

```
std.res <- pearson.res/sqrt(1 - hi) #standardize
adj.res <- std.res - mean(std.res) # mean-adjusted

# bootstrapping
B = 1000

beta.boot = beta.wild = matrix(0,p,B)

W <- diag(as.numeric(mu.hat))
G <- sqrt(W)%*%X

for(b in 1:B){

  eps.boot <- sample(adj.res,n,replace=TRUE) # residual bootstrap
  eps.wild <- std.res*rnorm(n) # wild bootstrap

  beta.boot[,b] = beta.hat + solve(t(G)%*%G)%*%t(G)%*%eps.boot
  beta.wild[,b] = beta.hat + solve(t(G)%*%G)%*%t(G)%*%eps.wild

}

# compute confidence intervals
# install.packages("glmtoolbox")
library(glmtoolbox)

int.boot = t(apply(beta.boot,1,quantile,probs=c(alpha/2,1-alpha/2)))
int.wild = t(apply(beta.wild,1,quantile,probs=c(alpha/2,1-alpha/2)))
int.prof = confint2(fit, level=1-alpha,test="lr")
int.wald = confint2(fit, level=1-alpha,test="wald")

column.names <- c("Lower bound","Upper bound")
row.names <- c("beta_1","beta_2","beta_3")
matrix.names <- c("Residual bootstrap","Wild bootstrap",
"Likelihood-ratio","Wald")

intervals <- array(c(int.boot,int.wild,int.prof,int.wald),dim=c(3,2,4),
dimnames = list(row.names,column.names, matrix.names))
intervals
, , Residual bootstrap

       Lower bound Upper bound
beta_1   1.4464596   1.5891100
beta_2   0.5913953   0.7497315
beta_3   0.8121010   0.9786808

, , Wild bootstrap

       Lower bound Upper bound
beta_1   1.4459307   1.5880854
beta_2   0.5937431   0.7418333
beta_3   0.8257604   0.9765748

, , Likelihood-ratio

       Lower bound Upper bound
beta_1   1.4452679   1.5919365
beta_2   0.5935432   0.7466844
beta_3   0.8177469   0.9757673

, , Wald

       Lower bound Upper bound
beta_1   1.4453670   1.5920353
beta_2   0.5936737   0.7468125
beta_3   0.8178775   0.9758967
```